

**UNITED STATES AIR FORCE**  
**SUMMER RESEARCH PROGRAM - 1991**

**SUMMER FACULTY RESEARCH PROGRAM**  
**(SFRP) REPORTS**

A248766

**VOLUME 4**

**ROME LABORATORY**  
**ARNOLD ENGINEERING DEVELOPMENT CENTER**  
**F. J. SEILER RESEARCH LABORATORY**

**RESEARCH & DEVELOPMENT LABORATORIES**

**5800 UPLANDER WAY**  
**CULVER CITY, CA 90230-8608**

**SUBMITTED TO:**

**LT. COL. CLAUDE CAVENDER**  
**PROGRAM MANAGER**

**AIR FORCE OFFICE OF SCIENTIFIC RESEARCH**

**BOLLING AIR FORCE BASE**

**WASHINGTON, D.C.**

Best Available Copy

**DECEMBER 1991**

**Best  
Available  
Copy**

AFOSR-TR-92-0170 (AFSC)  
This report has been reviewed and is  
being released IAW AFR 190-12  
unlimited.  
Program Manager

92-09041



92 4 08 004

REPORT DOCUMENTATION PAGE			Form 298-104-9 JMS No 1754-0188	
<small>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204 Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Project (0704-0188) Washington, DC 20503.</small>				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 9 January 1992	3. REPORT TYPE AND DATES COVERED 30 Sep 90-30 Sep 91		
4. TITLE AND SUBTITLE 1991 Summer Faculty Research Program (SFRP) Volumes 2-5b Vol. 4		5. FUNDING NUMBERS F49620-90-C-0076		
6. AUTHOR(S) Mr Gary Moore				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Research Development Laboratories (SDL) 5800 Uplander Way Culver City CA 90230-6608		8. PERFORMING ORGANIZATION REPORT NUMBER AFOSR-TR- 92 0170		
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/NI Bldg 410 Bolling AFB DC 20332-6448 Lt Col V. Claude Cavender		10. SPONSORING / MONITORING AGENCY REPORT NUMBER		
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT  UNLIMITED			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words)  <p>The purpose of this program is to develop the basis for continuing research of interest to the Air Force at the institution of the faculty member; to stimulate continuing relations among faculty members and professional peers in the Air Force; to enhance the research interests and capabilities of scientific and engineering educators; and to provide follow-on funding for research of particular promise that was started at an Air Force laboratory under the Summer Faculty Research Program.</p> <p>During the summer of 1991 170 university faculty conducted research at Air Force laboratories for a period of 10 weeks. Each participant provided a report of their research, and these reports are consolidated into this annual report.</p>				
14. SUBJECT TERMS			15. NUMBER OF PAGES	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED			18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED
			20. LIMITATION OF ABSTRACT UL	



**UNITED STATES AIR FORCE**  
**SUMMER RESEARCH PROGRAM -- 1991**  
**SUMMER FACULTY RESEARCH PROGRAM (SFRP) REPORTS**

**VOLUME 4**

**ROME LABORATORY**  
**ARNOLD ENGINEERING DEVELOPMENT CENTER**  
**F. J. SEILER RESEARCH LABORATORY**

**RESEARCH & DEVELOPMENT LABORATORIES**

**5800 Uplander Way**  
**Culver City, CA 90230-6608**

**Program Director, RDL**  
**Gary Moore**

**Program Manager, AFOSR**  
**Lt. Col. Claude Cavender**

**Program Manager, RDL**  
**Claude Baum**

**Program Administrator, RDL**  
**Gwendolyn Smith**

**Submitted to:**

**AIR FORCE OFFICE OF SCIENTIFIC RESEARCH**

**Bolling Air Force Base**

**Washington, D.C.**

**December 1991**



Accession No.	
NTIS GRA&I	
DTIC TAB	
Unannounced	
Justification	
By	
Distribution	
Ext	
A-1	

## **PREFACE**

Reports in this document are numbered consecutively beginning with number 1. Each report is paginated with the report number followed by consecutive page numbers, e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3.

This document is one of a set of 13 volumes describing the 1991 AFOSR Summer Research Program. The following volumes comprise the set:

<b><u>VOLUME</u></b>	<b><u>TITLE</u></b>
1	Program Management Report
<i>Summer Faculty Research Program (SFRP) Reports</i>	
2	Armstrong Laboratory, Wilford Hall Medical Center
3	Phillips Laboratory, Civil Engineering Laboratory
4	Rome Laboratory, Arnold Engineering Development Center, Frank J. Seiler Research Laboratory
5	Wright Laboratory
<i>Graduate Student Research Program (GSRP) Reports</i>	
6	Armstrong Laboratory, Wilford Hall Medical Center
7	Phillips Laboratory, Civil Engineering Laboratory
8	Rome Laboratory, Arnold Engineering Development Center, Frank J. Seiler Research Laboratory
9	Wright Laboratory
<i>High School Apprenticeship Program (HSAP) Reports</i>	
10	Armstrong Laboratory
11	Phillips Laboratory, Civil Engineering Laboratory
12	Rome Laboratory, Arnold Engineering Development Center
13	Wright Laboratory

## 1991 FACULTY RESEARCH REPORTS

### Rome Laboratory, Arnold Engineering Development Center, Frank J. Seiler Research Laboratory

<u>Report Number</u>	<u>Report Title</u>	<u>Author</u>
<u>Rome Laboratory</u>		
<b>Rome Air Development Center -- Griffiss (RADC)</b>		
1	Software Engineering Tools for Parallel Software Development	Dr. John Antonio
2	(Report Not Available at this Time)	Dr. Abdul Aziz Bhatti
3	A Taxonomy for Adaptive Fault Management in Survivable C <sup>3</sup> Systems	Dr. Rex Gantenbein
4	Analysis of the Electromigration Induced Failure in the VLSI Interconnection Components and the Multisection Interconnections	Dr. Ashok Goel
5	Optical Tunneling AND Gate/Test of Spatial Light Modulator/Binary Lens Light Beams	Dr. Philip Kornreich
6	Characterization of Radar Clutter as an SIRP	Dr. Jay Lee
7	Photocorrelation in Bi <sub>12</sub> SiO <sub>20</sub> And Semi-Insulating InP:Fe	Dr. Wallace Leigh
8	Development of a Methodology for Extracting Semantic Relations from Definitions	Dr. Elizabeth Liddy
9	Application-Based Utility Evaluation: A Discussion Leading to Assessing the Role of End-Users in the Exploitation of Virtual Reality Technology	Dr. Michael Nilan
10	Optical Fiber Amplifiers and Oscillators	Dr. Salahuddin Qazi
11	Collecting Data for Markov Models of Error Patterns on Data Communications Links	Dr. Wayne Smith
12	Approximating Neural Nets with C <sup>1</sup> Neural Nets	Dr. Michael Taylor
13	Optical Fiber Amplifiers and Oscillators	Dr. Kenneth Teegarden
14	Simulation Model Integration Methodology for Rome Laboratory	Dr. Jeffrey D. Tew
<b>Rome Air Development Center -- Hanscom (RADCH)</b>		
15	High Performance Microstrip Arrays for Polarimetric Bistatic Radar (PBR) Applications	Dr. Marat Davidovitz
16	User Assisted Information Extraction	Dr. Pradip Dey

**Report  
Number**

**Report Title**

**Author**

**Rome Laboratory (cont.)**

- |    |   |                       |
|----|---|-----------------------|
| 17 | Millimeter-Wave Noise Modeling Investigation  | Dr. Lawrence Dunleavy |
| 18 | Sinusoidal Transform Coder Parameter Manipulation Techniques and Their Use in Network and Data Storage Applications | Dr. Joseph Evans      |
| 19 | A General One-Dimensional III-V Heterojunction Device Simulator   | Dr. Ronnie Owens      |
| 20 | FDTD Analysis of the Radiation Properties of a Parabolic Cylinder Illuminated by a Very Short Pulse                 | Dr. Carey Rappaport   |

**Arnold Engineering Development Center (AEDC)**

- |    |  |                       |
|----|--|-----------------------|
| 21 | Thermodynamic and Transport Properties of Molecular Ions   | Dr. Murty Akundi      |
| 22 | Non-Intrusive Testing of Composite Aircraft Engine Components: I                                     | Dr. Laurence Jacobs   |
| 23 | Implementation of Multigrid in the PARC Code   | Dr. Steven McKay      |
| 24 | X-Ray Spectrometers for Pulsed Bremsstrahlung  | Dr. Carlyle Moore     |
| 25 | The Effect of Carbon Particle Combustion on the Infrared Signature Magnesium-Fluorocarbon Flare      | Dr. Olin Norton       |
| 26 | Software for 2D and 3D Mathematical Morphology   | Dr. Richard Peters II |
| 27 | A Review of CADDMAS  | Dr. Dean Smith        |
| 28 | Wake and Projectile Velocity Estimation/An Extended Kalman Filter Observer for an Altitude Test Cell | Dr. Mitchell Wilkes   |

**Frank J. Seiler Research Laboratory (FISRL)**

- |    |   |                       |
|----|---|-----------------------|
| 29 | New Reaction Transformation Using Nitronism Triflate  | Dr. Christopher Adams |
| 30 | Thermal Decomposition of TNT, NTO, and Their Mixtures via Isothermal Differential Scanning Calorimetry      | Dr. Gary Buckley      |
| 31 | Ternary Phase Diagram of MEIC/NaCl/AlCl <sub>3</sub>  | Dr. Do Ren Chang      |
| 32 | Photo-Electronic Nonlinear Excitations and Wave Propagation in Periodically Modulated Media                 | Dr. Marek Grabowski   |
| 33 | An Ab Initio Study of the Adducts between HF and HCl and Aluminum Hydrides, Halides, Hydroxides, and Oxides | Dr. Gilbert Mains     |

# **SOFTWARE ENGINEERING TOOLS FOR PARALLEL SOFTWARE DEVELOPMENT**

**John K. Antonio**  
School of Electrical Engineering  
Purdue University  
West Lafayette, IN 47907

**Richard C. Metzger**  
Rome Laboratory/COEE  
Griffiss AFB, NY 13441

## **Abstract**

The major objectives of this summer research project were to: (1) evaluate some of the currently available software engineering tools for developing and maintaining parallel software, (2) parallelize a computationally intensive portion of serial code contained within the Air Force's data fusion model and (3) propose new and/or enhanced tools and methodologies to aid in the development and maintenance of parallel software. Objectives (1) and (2) were accomplished concurrently by employing two software engineering tools to the process of parallelizing a small portion of the serial data fusion model. The primary focus here was not so much on the resulting parallel software, but rather on the *process* involved in developing the parallel code. The tools were found to be useful aids in the sense that they either simplified or automated certain low-level aspects of developing parallel software; thus allowing for rapid testing and prototyping of various system-level software designs. Future research in the area of parallel software engineering tools and techniques is suggested at the end of this report.

## **I. Introduction**

### ***A. Background***

Over the past several decades the state of software engineering for serial computers has matured into a quite sophisticated science. In contrast, the unique software engineering problems associated with developing and maintaining software for massively parallel computers have only recently begun to surface and be seriously addressed. While the concept of parallel computing is as old as the era of electronic computing, only within the past decade have commercially available and cost-effective parallel computing systems become available [6]. In February 1985 the Intel Corporation announced the iPSC, a computer based on a (then) new structure called the hypercube architecture [6]. Early demonstrations showed that for certain applications, a 128-processor hypercube could sustain a processing rate of 10 MFLOPS (million floating-point operations per second); roughly one-fifteenth the rate of the (then) fastest programs on the Cray X-MP-2X, at one-twentieth the cost. In 1986 Thinking Machines Corporation announced the Connection Machine, consisting of between 16 thousand and 64 thousand processors arranged on a very high speed network. While individually the nodes were not very powerful, collectively they could handle many symbolic computations at very high speeds [6,7].

Some parallel computers were built as far back as the 1960s, although most of these were more of a laboratory curiosity than a practical and cost-effective means of doing high-performance computing. Most notable of these early machines was the ILLIAC IV computer, constructed at the University of Illinois in the late 1960s. While limited memory and expensive hardware prevented proliferation, it inspired much interesting work in parallel algorithms. There have been many parallel machines built (both commercially and in research laboratories) since and prior to Intel's announcement of the iPSC. For a more comprehensive survey of the evolution of parallel computing, the reader is referred to the following references: [6,7,8].

### ***B. Motivation***

The following quote is taken from an editorial written by Peter J. Denning, past editor-in-chief of the *Communications of the ACM*, "... just beyond the hardware speed barrier lies another: the software barrier. We are good at understanding sequential algorithms, but we have little experience with algorithms that direct the activities of parallel processors. We do not know how to program the new [parallel] machines!" [6]. To date, most of the successes in obtaining significant speed-up by using parallel processing have been limited to fairly specific application areas such as signal processing, image processing, and

computation of certain classes of multivariable functions. There has been some progress in the area of parallelizing compilers (compilers which attempt to automatically parallelize conventional serial code), however, it is not clear if this approach alone will ultimately prove to be sufficient. Many researchers believe that a combination of several different approaches will eventually be required in order to develop efficient parallel software development techniques flexible enough for "general purpose" computing. The goals of this project were to try and tie together some currently available parallel software engineering tools, and suggest directions of future research.

### *C. Organization of the Report*

The report is organized in the following manner. In Section II, a brief description of the software engineering tools utilized and some background material on the data fusion problem are given. In Section III we report on the process of applying the software engineering tools to the task of parallelizing a small portion of the serial data fusion code. Section IV proposes some directions of future work. The Appendix contains some selected program listings.

## **II. Description of the Tools and the Data Fusion Problem**

### *A. Parallel Proto (PProto)*

According to its developers, "PProto is a rapid-prototyping environment that supports the software engineer in the definition of software system specifications and the prototyping and evaluation of specifications and designs" [5]. In the paragraphs below we briefly describe four of the basic components of PProto: (1) software simulation, (2) hardware simulation, (3) mapping software to hardware and (4) performance analysis. For a more detailed explanation, the reader is referred to the PProto User's Manual [5].

PProto uses both graphical and textual constructs in the description and simulation of the system software. At a high-level, the software is described graphically using nodes and links. The nodes represent processes and the interconnecting links represent the flow of control and/or data messages between the processes. Contained within each node is a description of its "behavior," which is written in a prototyping language called SSDL (System Specification and Design Language). Process nodes and links are created graphically by simply selecting the appropriate item from a graph editor menu. To edit the behavior of a process node (i.e., the SSDL program), the node is first selected followed by the selection of the behavior editor, which brings up a window for typing in the SSDL behavior for that node.

PProto also allows the user to create a model of the target hardware. For this, a hardware graph editor is provided which operates in a similar fashion to the graphical editor for creating the software model. In the hardware graph, the nodes represent processors and the links represent actual hardware busses. The relative speeds of the processors and busses can be assigned by selecting the appropriate item from a menu. The hardware editor also models memory modules, for which relative memory access times may be assigned. The hardware graph editor allows the user to quickly construct and edit the target hardware.

One of the very useful aspects of PProto is that it allows the user to view the system software independent from the hardware. So, once the software and hardware models are in place, the mapping editor is then used to map the processes (software) onto physical processors (hardware). By separating the hardware from the software, the user can (at least initially) focus on the correctness of the software. Later, by simulating various mappings onto a fixed hardware model, for example, the user can fine-tune the initial software design (if necessary) so as to achieve better overall performance and utilizations.

PProto provides various forms of output from the simulation which aid the software engineer in uncovering any bottlenecks (in both hardware and software) that may result from a particular mapping. The simulation itself is animated in the sense that the software process nodes change color (from blue to red) whenever a behavior within a process is executing. The interconnecting data/control links also change color when a message is transmitted across them. The hardware graph is animated as well with graphical performance meters placed beside the processors and busses showing the percentage of time each piece of hardware is used. The overall execution time (measured in simulation time units) is displayed as the simulation is running. At the end of the simulation a summary text file is printed to standard output.

### *B. Computing Surface Tools (CSTools)*

CSTools is a collection of support programs supplied by Meiko Scientific Inc. for use with their transputer based parallel processing systems. The transputer system used was the MK202, which consists of a bank of sixteen T800 transputers (in the computational engine) hosted by a Sun 4/470 workstation. The unique property of the transputer system is that the interconnection topology (between the transputer nodes) can be controlled via software commands, provided that the in- and out-degree at each node does not exceed four. The term *computing surface* refers to the set of transputer nodes and interconnecting hardware busses used by a particular application. One of the tools (within CSTools) enables the user to either explicitly define the interconnection topology or allow a generic



topology to be chosen by the CStools operating system. Examples of common interconnection topologies are the binary-tree, the four-neighbor mesh and the linear array. There is also a tool which enables the user to download parallel programs from the host machine onto the computing surface nodes. We point out here that neither of these tools are graphically based and are, therefore, a bit cumbersome to use. Better user interfaces for these tools would be graphical ones like those implemented in PProto for describing software, hardware, and the mapping between them. This issue will be addressed in a latter section on future directions of research.

There are also tools within CStools to aid the programmer in setting-up communication protocols for exchanging messages between processors. The programming environment associated with these communication protocols uses the concept of a *computing surface network (CSN)*, which allows the programmer to abstract away from the physical interconnection topology associated with a particular configuration of the computing surface. From the programmer's point of view, messages are sent from an origin node to a destination node via a transport through the CSN. The CSN is responsible for the actual store-and-forward routing associated with passing the messages through the physical interconnection topology.

CStools supports four types of protocols for interprocessor communication: (1) blocking-synchronous, (2) blocking-asynchronous, (3) nonblocking-synchronous and (4) nonblocking-asynchronous. In the blocking protocols, control is not returned to the caller process (i.e., the sender of the message) until "completion" of the transmission (thus, execution at the caller process is *blocked* until an acknowledgement is received). For the synchronous cases, "completion" is defined whenever the message is successfully received at the destination node, while in the asynchronous cases "completion" is defined when the message successfully enters the CSN. In the nonblocking cases, control is given back to calling process immediately.

### *C. The Data Fusion Problem*

Data fusion encompasses a very broad range of problems and is therefore difficult to define precisely. However, within the domain of Air Force applications, data fusion usually refers to those settings in which vast amounts of sensor data (collected from remote locations) is sent to a central location for processing and interpretation. The central processing location is referred to as the *fusion center*. The underlying objective of the fusion center is to "make sense" of the incoming data in a real-time or *near* real-time fashion. Depending on the specific application, the definition of "making sense" of the data varies. The data fusion

model studied in this project primarily involves filtering and correlating incoming data from remote electronic sensors. The data reported by the remote sensors provide the fusion center with a piece of intelligence information about various military targets (such as target position and/or velocity estimates). For this project we focused on the Advanced Sensor Exploitation (ASE) model, which is a current data fusion model within the Air Force. A functional overview of the ASE model is shown in Fig. 1.

Data fusion problems are inherently computationally intensive for several reasons. First, due to the massive amount of incoming data to be processed in a typical scenario, even the simplest data fusion algorithms may suffer from slow execution on conventional serial computers. Second, the accuracy of much of the incoming data may be very low, thus the fusion algorithms must be sophisticated enough to make intelligent decisions based on noisy measurements. Finally, not only is there a variance in the accuracy of the data, but portions of incoming data records may be missing or incomplete.

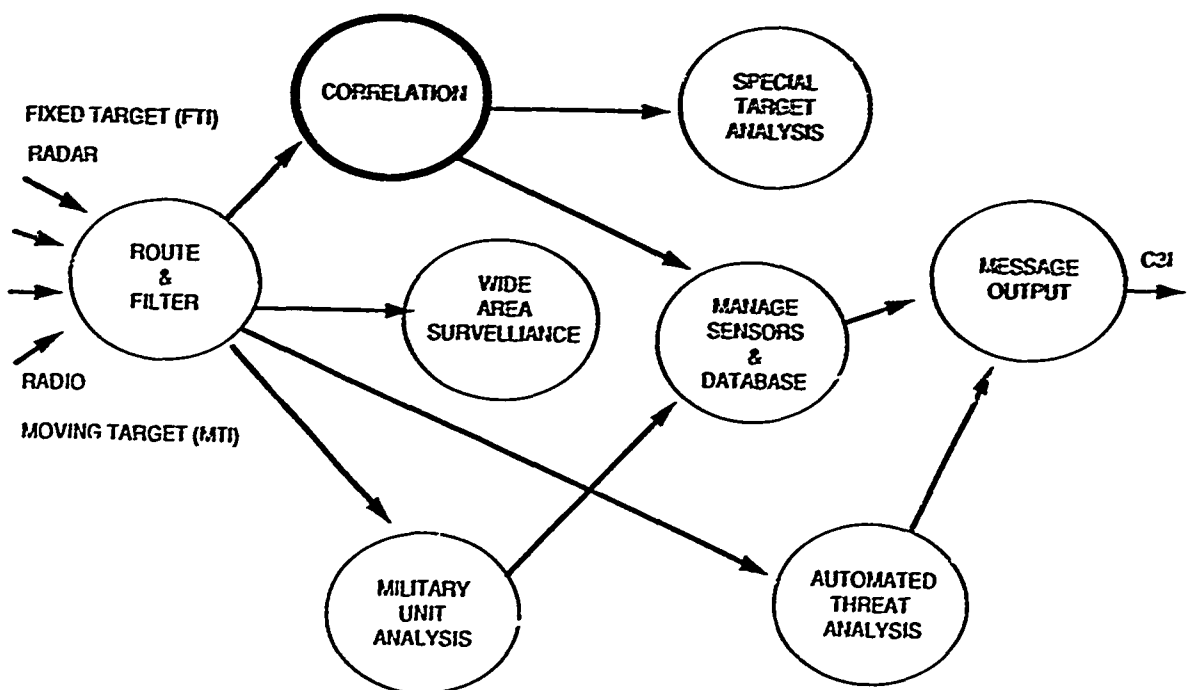


Fig. 1: A functional overview of the ASE data fusion model.

### III. The Parallelization Process

In this section we describe the process taken to parallelize a small portion of the data fusion program. It should be understood that the approach taken is by no means the only possibility, and we do not claim that the overall approach taken is the "best" possible

solution (as measured by any particular software engineering metric). Our underlying objective was to employ an approach in which the available tools could be evaluated. As of the writing of this report, some of the latter stages of the process were still on-going, however, enough progress was made (or experience gained) to report on all stages.

#### *A. High-Level Understanding of the Functionality*

The current data fusion model contains nearly 500,000 lines of serial FORTRAN code. To make our project meaningful, we needed to extract a portion of the code which was both parallelizable and manageable. Because we did not have an automated tool for detecting parallelism contained within serial FORTRAN code, we resorted to the task of learning some of the high-level functionality of the data fusion model. After approximately two weeks of reading documentation and talking with people directly involved in the original development of the model, the correlation node was chosen as a prime candidate for parallelization. The correlation node was computationally intensive and believed to be highly parallelizable. Within the correlation node there is a module called entity association. The purpose of entity association is to determine whether a new (incoming) target report "associates" with known targets stored in the data base. If it does, then a confidence measure is updated within the report record stored in the data base. On the other hand, if the new target report does not associate with any of the reports in the data base, then it is stored as a new "candidate" target record in the data base.

The underlying measure of association (MOA)—which is used to determine the "closeness" between two target reports—is derived from a statistically based measure called the Mahalanobian metric. The MOA value is a real number between zero and one. An MOA value of zero implies that two reports are not associated, while a value of one represents the highest possible degree of association. The computation of the MOA between two reports requires a position vector and covariance matrix from each report. The mathematical formula for computing the MOA is given by

$$\text{MOA} = \exp [1/2 (x_t - x_c)^T R^{-1} (x_t - x_c)],$$

where  $x_t$  and  $x_c$  denote position vectors of a target report (from the sensors) and a candidate report (from the data base), respectively. The variable  $R$  is a matrix defined by

$$R = R_t + R_c,$$

where  $R_t$  and  $R_c$  are the covariance matrices for the target and candidate reports, respectively. The covariance matrices capture the uncertainty inherent in the position data as measured by various types of sensors. More detail on the theory behind the Mahalanobian metric can be found in reference [1] and the references therein.

### ***B. Serial Software Mock-Up***

Within the entity association module of the data fusion model, there is a section of code which determines the maximum MOA value between an incoming target report and a collection of candidate reports stored in the fusion center's data base. Our starting point for this project was to assume that the collection of candidate reports had been selected. Thus, the portion of code of interest was that part which finds the maximum value of MOA. The structure of the serial code for finding the maximum MOA is shown in Fig. 2.

```
begin maximum MOA calculation
  do for each target
    do for each candidate
      compute MOA
      if MOA > current_max_MOA then
        current_max_MOA := MOA
      end if
    end do
    update data base
  end do
end maximum MOA calculation
```

Fig. 2: The structure of the serial code for finding maximum MOA values.

We wrote three FORTRAN programs for emulating the above Maximum MOA part of the serial model. Two of the programs simply create synthetic target and candidate data files. The third program actually computes maximum MOA values between target and candidate data records. An overview of this serial system is shown in Fig. 3. Due to space limitations, we shall not describe the details of the actual serial FORTRAN programs nor the data structures used to store the target records, however, some selected program listings and output files can be found in the appendix. The serial machine used was a VAX/780, as this machine was also the target machine used to execute the actual ASE model.

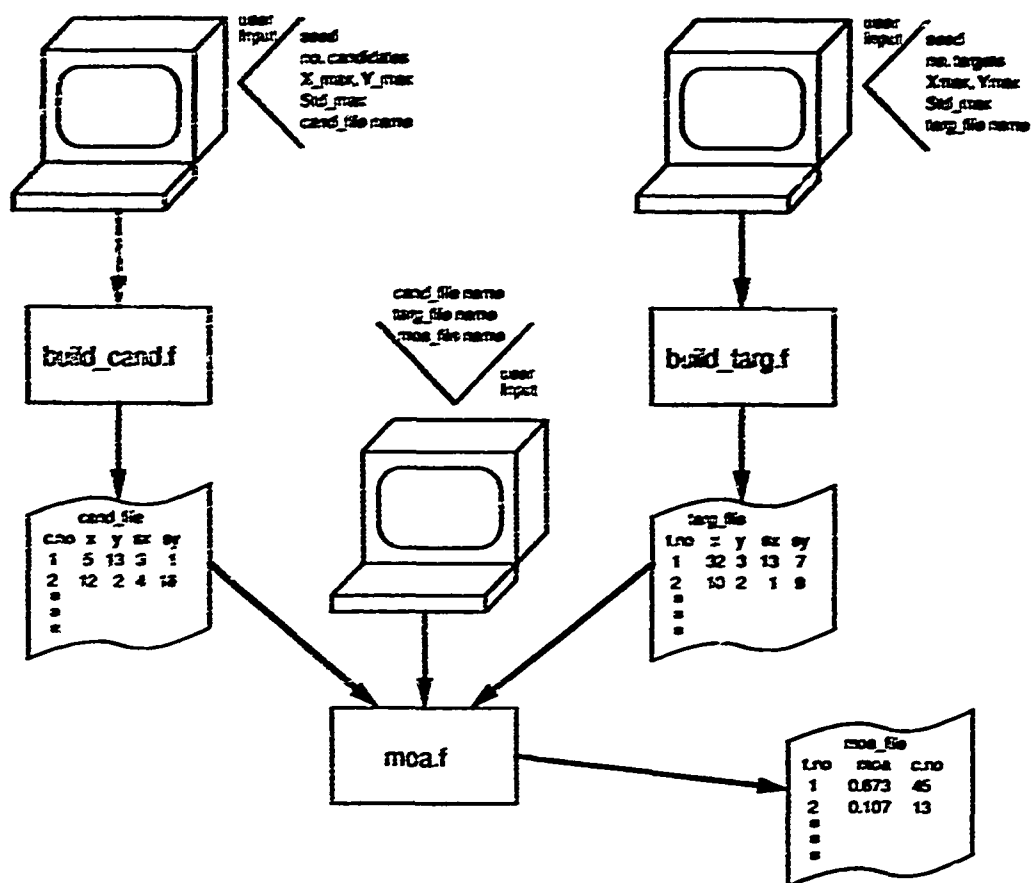


Fig. 3: Overview of the serial software mock-up.

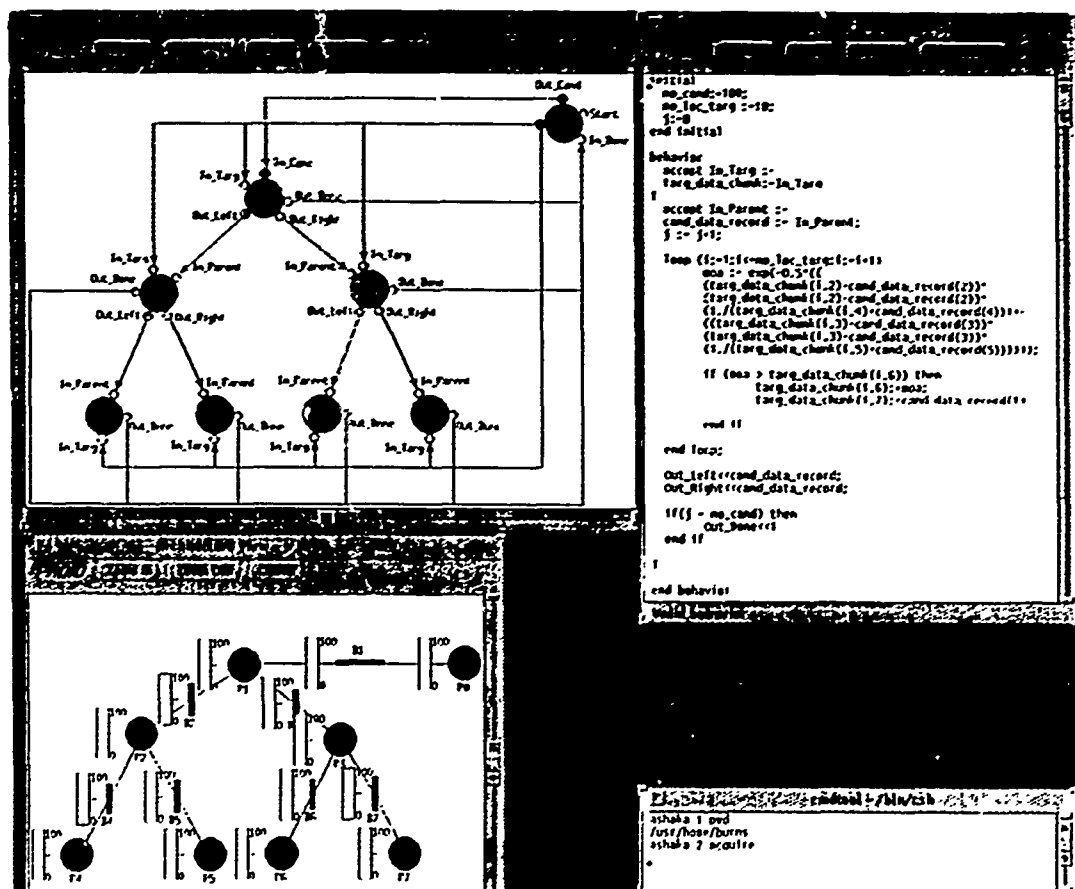
### C. The Basic Parallelization Strategy

Because our goal was to ultimately implement the parallel software on the transputer system, the following strategy was developed under the assumption that the target machine would have a distributed memory architecture.

The simple looping structure associated with the serial code (for computing maximum MOA values) made the task of deciding on a basic parallelization strategy fairly straightforward. We decided on the following approach. (1) Partition the target records evenly across the nodes (i.e., store equal-sized chunks of target data in each local memory module). (2) Pipeline candidate records through the network. (3) Whenever a node receives a candidate record, it updates its current maximum MCA values (one for each target record) and forwards the candidate record to the appropriate neighbor(s) in the broadcast tree.

### ***F. Parallel Software System Design (PProto)***

We used PProto to evaluate the merit of our basic parallelization strategy at a system software level. Two distinct implementations of the basic strategy were considered. In the first implementation the processors (and processes) were connected as a binary tree and in the second they were connected as a linear array. Figs 4 and 5 show the graphical views of these two implementations in both hardware and software. Note that the "lone" nodes in each figure represent the host machine and the other nodes denote the transputers. In the software graphs note that there are many more links than in the hardware graphs, because the host node must have a *logical* connection from itself to each other node. These logical connections are used to send the chunks of target data from the host down to each transputer. Also, there are logical links from each transputer node to the host which are used by the transputers to send their final results to the host. In the hardware graphs the links represent physical busses between the actual processors.



**Fig. 4: Graphical view of the binary tree implementation on PProto.**



logarithm of the number of processors while the filling time for the linear array grows linearly with the number of processors. With only seven parallel processors, this asymptotic effect was not a dominant factor.

#### *E. Parallel Software Implementation (CSTools)*

This portion of the parallelization procedure was not completed by the end of the project period. However, we did gain experience with programming the transputers for a different application, and therefore we can briefly explain the expected transition from the system software design to actual implementation on the transputers. First, the hardware can be configured as either the binary tree or linear array using the appropriate tool from CSTools. Second, the computational part of the behaviors from PProto can be converted to either C or FORTRAN easily (requires simple looping, math operations, and comparisons). Finally, the communication portions of the behaviors should also be fairly straightforward to implement with the CSTools protocols. In particular, the transmitting of data out through a port in the SSDL behaviors is accomplished by a blocking transmission through a transport, as defined under the CSTools programming environment.

### **IV Directions of Future Work**

#### *A. A Suggested Enhancement to PProto.*

From our experience with PProto, it appears that the delay associated with sending a message across a communication bus is independent of the size of message being sent. In our implementations of the maximum MOA calculation, we assigned a relative speed of unity to the busses and discovered that only one time unit was charged for sending either a chunk of 10 target records or a single candidate record. While PProto is a system software tool, we believe that a finer level of detail should be incorporated in estimating the communication delays.

#### *B. A Suggested Enhancement to CSTools*

The current tools within CSTools for defining the interconnection topology and mapping the processes onto processors require the user to type-in the correct commands. These commands are quite detailed and can become confusing when the number of processors used exceed 3 or 4. We recommend that a graphical method be implemented based on the same techniques developed by the designer of PProto. So, the user would have a graphical view of the hardware and software and a mapping editor for assigning processes to processors. The implementations of this suggestion would involve merging a portion of



the graphical technology developed for PProto with the current CStools.

### *C. A Tool for Balancing Communication and Computation in Parallel Systems*

There seems to be a gap in the availability of tools or techniques for aiding the software engineer in striking a balance between the computational and communicational aspects of developing parallel software. Through our own experiences in parallel programming, we have found that if the ratio of communication to computation is too high, then substantial bottlenecks may arise which can degrade overall performance significantly.

At the higher levels of developing parallel software, the software engineer may not be too concerned with the communication requirements and overhead of the software system. Instead, issues of correctness and partitioning of code and data are usually considered. Eventually, however, when the parallel program is implemented on a particular piece of hardware, the software engineer may find severe bottlenecks caused by contention for the communication resources. For example, if a particular software process transmits an average of 100 bytes of data per every 1 msec of computation (an average of 100 Kbytes/sec) across a communication channel having a capacity of only 10 Kbytes/sec, then obviously queues will eventually overflow and potentially cause bottlenecks and/or deadlocks.

A question now arises. Should the software engineer be responsible for counting clock cycles of computation between communications to ensure that the software executes efficiently on the target hardware? We believe the answer is no for several reasons. First, such an approach would force the software engineer to "commit" to a particular parallel machine, and therefore design the software around the hardware specifications. Second, such an approach is simply too tedious. Finally, such an approach does not lend itself to adapting the software to execute on different machines.

We propose to develop a tool to aid the software engineer in controlling the flow of transmissions so as not to overload the communication network. The benefits of employing a flow control mechanism are well understood in the areas of telecommunication/data networks and automobile traffic control. For instance, in telecommunication networks, certain callers may be denied network access whenever the utilization of communication resources exceeds a certain threshold. If properly controlled, this throttling of network access will result in better overall throughput. Similarly, the proper use of traffic lights can have the same effect by controlling the flow of traffic on city streets. Our approach is to employ similar techniques in the control of traffic between parallel processors.

## References

- [1] S. D. Allen, *et al.*, "Final Technical Report: ASE Implementation," PAR Technology Corporation, Contract No. F30602-80-C-0298, Data Item: B006, 28 October, 1983.
- [2] J. F. LoSecco, *et al.*, "Final Technical Report: Target Recognition for Electronic Combat," Synectics Corporation, Contract No. F30602- 87-C-0150, Data Item: A008, 8 December, 1989.
- [3] E. Waltz and J. Llinas, *Multisensor Data Fusion*, Artech House: Boston, MA, 1990.
- [4] A. Evans and N. Coulson, *CSTools Tutorial for C Programmers*, Meiko Limited, 1990.
- [5] *PProto User's Manual*, International Software Systems, Inc., Contract No. F30602-89-C-0129, 14 June, 1991.
- [6] P. J. Denning, "Parallel Computing and Its Evolution," *Communications of the ACM*, Vol. 29, No. 12, December 1986.
- [7] K. A. Frenkel, "Evaluating Two Massively Parallel Machines," *Communications of the ACM*, Vol. 29, No. 8, August 1986.
- [8] S. E. Miller, "Final Technical Report: A Survey of Parallel Computing (Second Edition)," Amherst Systems, Inc., Contract No. F30602-87-C-0082, 1 September, 1988.

**Appendix**  
**Selected Program Listings and Outputs**

```

*****
*
*   Program File Name: MOA.FOR
*   Programmer: John Antonio
*   Date: June 19, 1991
*
*   Description: This program computes the maximum measure of
*                 association (MOA) between a set of known candidates
*                 and a set of target reports. For each target report,
*                 the candidate having the maximum MOA is determined and
*                 written to an output file. Each candidate and target
*                 record consists of five items: an identification number,
*                 an x-coordinate, a y-coordinate, a standard deviation of
*                 error in the x direction, and a standard deviation of error
*                 in the y-direction. Each record of the output file
*                 consists of three items: the identification number of
*                 the target, the maximum MOA value, the identification
*                 number of the candidate associated with the maximum
*                 MOA value. Candidate data files are created by the
*                 program BUILD_CAND.FOR, and target data files are
*                 created by the program BUILD_TARG.FOR.
*
*****

implicit          none
integer*4         no_cand, no_targ, i, j, dummy
integer*4         maxmoa_index(10000)
character*16      cand_file, targ_file, out_file
real*4            xt_coord(10000), yt_coord(10000)
real*4            xt_std(10000), yt_std(10000)
real*4            xc_coord(10000), yc_coord(10000)
real*4            xc_std(10000), yc_std(10000)
real*4            moa(10000), maxmoa(10000)
real*4            t1, t2, delta1, delta2

*
*****
*   Input data file names from the user.
*
*****

1      write(5,1)
      format(' Input name of candidate input file')
      read(5,6) cand_file
      write(5,2)
2      format(' Input name of target input file')
      read(5,6) targ_file
      write(5,3)
3      format(' Input name of association output file (to be created)')
      read(5,6) out_file
6      format(a)
*
*****
*   Start the "read/compute/write" timer
*
*****

      t2 = secnds(0.0)

*
*****
*   Open the input data files
*   Note: The file names are chosen by the user
*
*****

      open(unit=9, file=cand_file, status='old')
      open(unit=10, file=targ_file, status='old')

*
*****
*   Read in the number of candidates and targets
*
*****

```

```

*****
*
      read(9,*)
      read(9,*)
      read(9,99) no_cand
      read(9,*)
      read(9,*)
      read(10,*)
      read(10,*)
      read(10,99)no_targ
      read(10,*)
      read(10,*)
99      format(10x, i5)
*
*****
*      Read in the x,y coordinates and standard deviations from the      *
*      candidate and target data files.                                *
*****
*
      do i=1, no_cand
        read(9,*) dummy,xc_coord(i),yc_coord(i),xc_std(i),yc_std(i)
      end do
      do j=1, no_targ
        read(10,*)dummy,xt_coord(j),yt_coord(j),xt_std(j),yt_std(j)
      end do
      close(9)
      close(10)
*
*****
*      Start the computation timer                                    *
*****
*
      t1=secnds(0.0)
*
*****
*      Start the MOA calculation                                    *
*****
*
      do j=1, no_targ
        do i=1, no_cand
          moa(i)=exp(-0.5*((xt_coord(j)-
&          xc_coord(i))**2)*
&          (1./ (xt_std(j)+xc_std(i)))+
&          ((yt_coord(j)-yc_coord(i))**2)*
&          (1./ (yt_std(j)+yc_std(i)))))
        end do
        maxmoa(j)=0.0
        maxmoa_index(j)=0
        do i=1,no_cand
          if(moa(i).gt.maxmoa(j)) then
            maxmoa(j)=moa(i)
            maxmoa_index(j)=i
          end if
        end do
      end do
*
*****
*      End of the MOA calculation                                    *
*****
*
      stop computation timer
*****
*
      delta1=secnds(t1)
*

```

```

*****
*      Open the output file and write out the results.      *
*****
*
      open(unit=11, file=out_file, status='new')
*
      write(11,77) out_file
77      format(' File name: ', a)
      write(11,78)
78      format(' This data file was created by moa.for')
      write(11,79) cand_file, targ_file
79      format(' Candidate input file: ', a, 'Targ. input file: ',a)
      write(11,799) no_cand, no_targ
799      format(' No. of candidates: ',i5,11x,' No. of targets: ',i5)
      write(11,80) delta1
80      format(' Time to compute all of the maximum MOAs: ', f10.5)
      write(11,*)
      write(11,81)
81      format(' Targ. No.    Max. MOA    Max MOA Cand. No.')

      do j=1, no_targ
         write(11,111) j, maxmoa(j), maxmoa_index(j)
      end do
111      format(i5, 9x, f7.5, 8x, i5)
*
*****
*      stop "read/compute/write" timer      *
*****
*
      delta2=secnds(t2)
*
      write(11,*)
      write(11,82) delta2
82      format(' Time for reading, computing, and writing: ',f10.5)
      write(5,*) delta1, delta2
*
      close(11)
*
      stop
      end

```

\*\*\*\*\*

File name: moa22\_780.dat

This data file was created by moa.for

Candidate input file: c2.dat

Targ. input file: t2.dat

No. of candidates: 100

No. of targets: 100

Time to compute all of the maximum MOAs: 1.65234

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.69514	64
2	0.96774	91
3	0.86108	35
98	0.95123	59
99	0.10026	14
100	0.73544	89

Time for reading, computing, and writing: 4.81250

\*\*\*\*\*

File name: moa23\_780.dat

This data file was created by moa.for

Candidate input file: c2.dat

Targ. input file: t3.dat

No. of candidates: 100

No. of targets: 1000

Time to compute all of the maximum MOAs: 9.12891

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.41403	81
2	0.00127	94
3	0.61967	70
998	0.91219	100
999	0.62772	18
1000	0.84568	15

Time for reading, computing, and writing: 18.22656

\*\*\*\*\*

File name: moa24\_780.dat

This data file was created by moa.for

Candidate input file: c2.dat

Targ. input file: t4.dat

No. of candidates: 100

No. of targets: 10000

Time to compute all of the maximum MOAs: 83.86719

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.78472	81
2	0.66918	97
3	0.84246	55
9998	0.97935	48
9999	0.78605	28
10000	0.96162	17

Time for reading, computing, and writing: 158.89063

\*\*\*\*\*

\*\*\*\*\*

File name: moa32\_780.dat

This data file was created by moa.for

Candidate input file: c3.dat

Targ. input file: t2.dat

No. of candidates: 1000

No. of targets: 100

Time to compute all of the maximum MOAs: 6.52734

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.95123	715
2	0.98347	907
3	0.89484	593
98	0.95919	787
99	0.71405	360
100	0.86688	957

Time for reading, computing, and writing: 13.91016

\*\*\*\*\*

File name: moa33\_780.dat

This data file was created by moa.for

Candidate input file: c3.dat

Targ. input file: t3.dat

No. of candidates: 1000

No. of targets: 1000

Time to compute all of the maximum MOAs: 86.08984

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.93359	489
2	0.78600	637
3	0.95123	999
998	0.95230	375
999	0.91952	978
1000	0.92609	296

Time for reading, computing, and writing: 97.34766

\*\*\*\*\*

File name: moa34\_780.dat

This data file was created by moa.for

Candidate input file: c3.dat

Targ. input file: t4.dat

No. of candidates: 1000

No. of targets: 10000

Time to compute all of the maximum MOAs: 853.51953

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.96882	385
2	0.89593	678
3	0.92169	6
9998	0.96813	870
9999	0.90484	319
10000	0.91034	19

Time for reading, computing, and writing: 924.83203

\*\*\*\*\*



\*\*\*\*\*

File name: moa42\_780.dat

This data file was created by moa.for

Candidate input file: c4.dat      Targ. input file: t2.dat

No. of candidates: 10000      No. of targets: 100  
Time to compute all of the maximum MOAs: 92.02734

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.98658	7865
2	0.98400	7562
3	1.00000	6209
98	0.97102	4928
99	1.00000	5451
100	1.00000	4650

Time for reading, computing, and writing: 142.35938

\*\*\*\*\*

File name: moa43\_780.dat

This data file was created by moa.for

Candidate input file: c4.dat      Targ. input file: c3.dat

No. of candidates: 10000      No. of targets: 1000  
Time to compute all of the maximum MOAs: 974.14063

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.97403	4731
2	1.00000	2160
3	1.00000	7554
998	0.96722	8505
999	0.96227	6606
1000	1.00000	4341

Time for reading, computing, and writing: 1021.78906

\*\*\*\*\*

File name: moa44\_780.dat

This data file was created by moa.for

Candidate input file: c4.dat      Targ. input file: t4.dat

No. of candidates: 10000      No. of targets: 10000  
Time to compute all of the maximum MOAs: 8873.75000

Targ. No.	Max. MOA	Max MOA Cand. No.
1	0.98621	7913
2	1.00000	3057
3	1.00000	1586
9998	1.00000	4948
9999	0.97531	5186
10000	1.00000	1838

Time for reading, computing, and writing: 8979.98047

**Dr. ABDUL A. BHATTI**

**REPORT NOT AVAILABLE**

**AT TIME OF  
PUBLICATION**

# A TAXONOMY FOR ADAPTIVE FAULT MANAGEMENT IN SURVIVABLE C3 SYSTEMS

Final Report  
AFOSR Summer Faculty Research Program  
16 August 1991

Rex E. Gantenbein, Associate Professor  
Department of Computer Science  
University of Wyoming  
P.O. Box 3682  
Laramie WY 82071

## ABSTRACT

Most strategies for fault management are effective for a narrow range of fault classes. In survivable C3 systems, a wide range of operating environments may be encountered that require different strategies to be used at different times. This report presents a classification for fault management strategies that can be used to define the most appropriate methods for assuring survivability under conditions that can suddenly and drastically change.

Two metrics by which this adaptivity can be specified and evaluated -- the *objective function* and a *consistency measure* -- are defined. These metrics can be used to determine how well a survivable system meets its requirements under the current operating environment. Policies are discussed that group strategies by the tradeoffs that can be made when the system is unable to meet its requirements for objectives or consistency due to failures, as well as by the fault classes that can be handled.

# A TAXONOMY FOR ADAPTIVE FAULT MANAGEMENT IN SURVIVABLE C3 SYSTEMS

Final Report  
AFOSR Summer Faculty Research Program  
16 August 1991

Rex E. Gantenbein, Associate Professor  
Department of Computer Science  
University of Wyoming  
P.O. Box 3682  
Laramie WY 82071

## INTRODUCTION

In the context of computing systems, *dependable* has been defined as a qualitative term that characterizes a system that can be justifiably trusted to deliver the required service when needed [Avizienis89]. Such systems have become important of late, due to the increased dependency of commercial and military operations on complex computations that cannot be implemented as simple, easily testable sequential programs. As these systems are used in more complex environments, the range of circumstances that are encountered in the course of their operation increases. A complex system intended to be dependable must be designed to meet and handle a wide variety of potential problems in the course of its operation.

Unfortunately, most strategies intended to provide dependability consider only a few very specific issues. Even the notion of dependability itself is typically divided into subsections, such as reliability, availability, performance, etc., each of which is addressed by a number of methods that claim to increase the system's ability to withstand perturbations in its operating environment. These methods provide increases in the system's resilience to particular types of problems, but the assumptions that are made about the environment (often to make the proposed solution tractable or efficient) limit the effectiveness of the method to exactly those cir-

cumstances specified in the assumptions. As long as the environmental assumptions are not violated, a dependable system should be able to achieve its specified goals, even in the face of definable failures such as processor or communication losses.

The result of this optimization for a particular operating environment is that most dependable systems tend to be rigid and brittle. Any violation of a system's assumptions may cause it to behave in ways that are, at best, suboptimal and, at worst, unexpected. Normally, the system designer attempts to specify the expected environment as part of the requirements definition, which sets out the expectations for the system in terms of things such as functionality, performance, and precision of produced results. Since specification is still an imperfect craft, complete description of an environment is difficult. Many assumptions about the operating environment of a system are made implicitly by the designer and implementor.

Since the same strategy that works under one set of assumptions about the operating environment may not work under a different set, what was an effective strategy in one case may be ineffective and even harmful under another. A rapid and drastic change in the environment, such as is sometimes seen in field-deployed C3 systems under combat conditions, can cause catastrophic failure in the system if the highly adverse environment is not included in the design requirements. On the other hand, a system that is capable of surviving in a combat environment may not be effective or efficient in "normal" (noncombat) usage. Even the mission of a C3 system can change over time (from strategic planning to theater operation to tactical battle management, for example), and strategies appropriate for one mission may not be for another.

There are numerous factors affecting fault management that can suddenly change. For example, one factor common to many approaches is the ability to replace system components

that have malfunctioned. Obviously, any system in which a component cannot be replaced is more susceptible to additional malfunctions than one in which replacement is possible. Since some strategies for avoiding failure are more effective for single-component malfunctions than for multiple malfunctions, the designer must choose a strategy appropriate for what he or she believes to be the most likely (or perhaps the most dangerous) circumstances.

It is easy to see that in this case the same system can encounter widely different circumstances during its operation. The system may at any time find that replacements have suddenly become unavailable. One could also imagine that the system could be deployed in a mode that precluded the replacement of peripherals except during times of scheduled maintenance. Under these conditions, failure of a component would be very detrimental to the potential for continued operation of the system. In fact, it might be useful to consider replacing components before they fail, rather than after, so that their loss is less likely. Predicting the likelihood of failure in a component is a difficult task, but it may be necessary if the system cannot be taken down for an extended period of time, as when an threat is imminent, for example.

Clearly, what this system needs is to be *adaptive*, that is, able to operate under different sets of environmental conditions without violating its requirements, which may also vary with the conditions. In a *survivable* system, the primary requirement is to avoid catastrophic failure, even if it means relaxing or ignoring requirements that cannot be met under adverse conditions. This behavior is usually referred to as *graceful degradation*. Since the transition between one environment and another may occur too quickly for an operator to detect, or may be due to problems within the system itself not observable by an operator until it is too late, the system must also be able to detect (or predict) such transitions in order to be able to con-

trol (or prevent) the failures that can accompany them.

## ADAPTIVE FAULT MANAGEMENT IN SURVIVABLE SYSTEMS

The purpose of this paper is to present a categorization of strategies for *fault management* in survivable systems that defines the circumstances under which strategies may be appropriately applied. Fault management is the application of techniques in a computing system to prevent the system from failing to deliver its required service due to faults in any of its components (hardware or software). We use the term "fault management" rather than "fault tolerance" because the strategies used to avoid catastrophic failure include prediction and approximation as well as the more well-known methods such as masking and recovery. Moreover, although not addressed explicitly in this paper, it is a commonly held belief that the achievement of highly dependable systems requires the application of fault management strategies at all stages of a system's life, from design to testing to operation to maintenance. The categorization to be presented here may be applied equally well to strategies for design and maintenance and to operational strategies.

We address these strategies in the context of distributed C3 systems. For the greatest likelihood of survival, a C3 system must be distributed over several processing nodes, geographically dispersed and loosely coupled by communication links. We are particularly interested in heterogeneous nodes that are devoted to a single application, but capable of autonomously operating in both dedicated and general-purpose modes. These processors should support reconfiguration and redistribution of the system components in case of failures in any of the three basic classes of system resources [Kohler81]: processing (generating or transforming information), data (storing information), or communication (transmitting information).

*Failure* is the inability of a system to deliver a required service. The cause of a failure is a malfunction in the system operation that results in an incorrect system "state" which renders the system incapable of delivering the service. This incorrect state is an *error*. The cause of an error is a *fault* in the software or hardware that was expected to produce the service. This hierarchical definition can be extended for components so that we can say a failure in a component becomes a fault in a higher-level component that depends on the undelivered service.

A system that supports adaptivity can potentially achieve greater resilience to failures under (possibly rapidly) changing environments by allowing the system to dynamically change its fault management strategy. *Predictive* strategies are those that stress avoidance of errors, based on what the system thinks will happen, while *reactive* strategies are those that stress tolerance of errors that have already occurred and been detected. While it is impossible for a system to control its environment (or external state) and therefore the results of its execution (or internal state), it is possible to predict or react to changes in those states and apply different strategies based on those predictions or reactions.

We therefore define *adaptive fault management* as the dynamic selection of fault management strategies and the mechanisms by which the strategies are implemented, based on predictive or reactive information about the system state, either external or internal. We also define a general model of adaptive fault management, as shown in Figure 1. Given a library of fault management strategies, the system must evaluate the current external state (including operation mode and mission) and the current internal state (including error rate, fault type, resources available, resource utilization, and performance metrics) to determine the current computing environment. Using a set of rules that relate the environment to the requirements



for the system, an adaptive behavior manager selects a fault management strategy. This strategy is used in scheduling of system tasks and fault management mechanisms and in determining the parameters for the execution of those mechanisms.

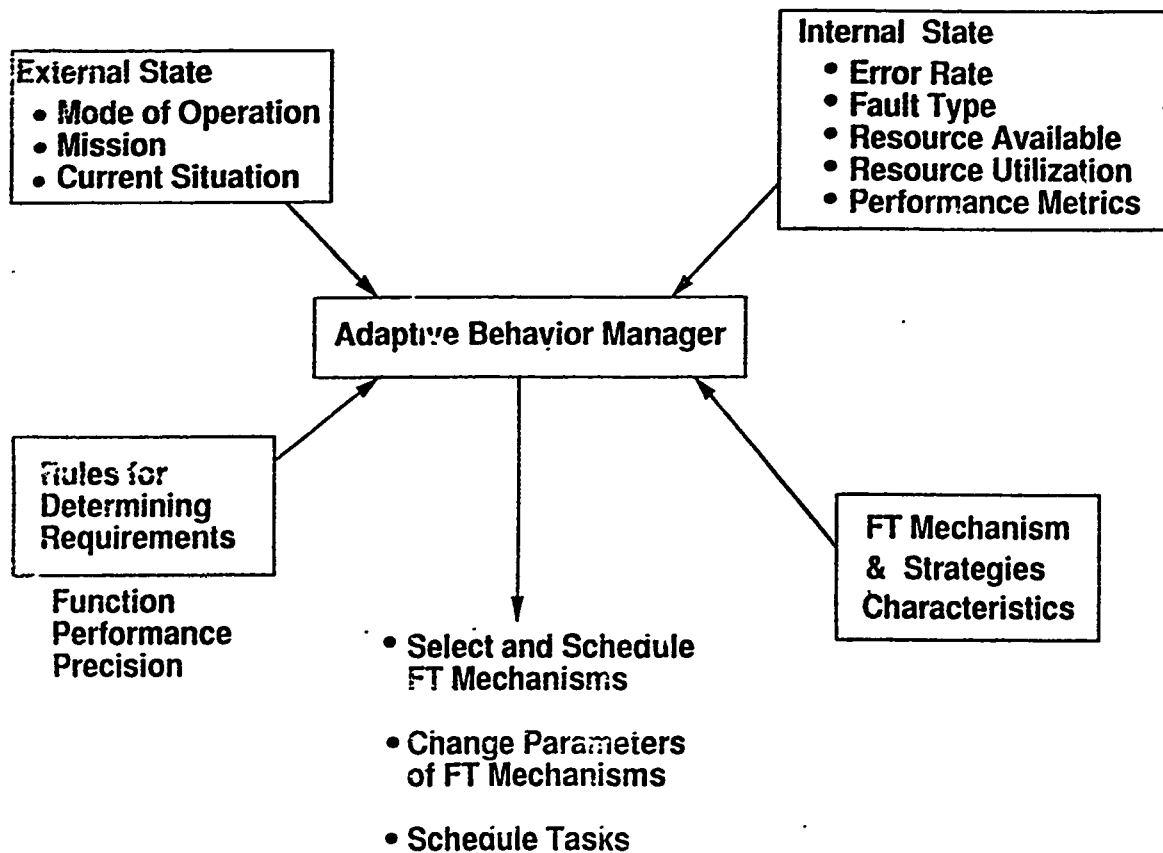


Figure 1. A model for adaptive fault management.

## SPECIFYING ADAPTIVITY IN SOFTWARE SYSTEMS

In order for adaptivity such as we are describing here to be effectively used, it is necessary that the system designer be able to specify the acceptable system behavior in a given environment. We need a way by which the system can be evaluated within an environment to see whether the current fault management strategy is still effective or whether it needs to be changed. In this section, we define two metrics that can be used to define a system's require-

ments for adaptivity and to evaluate how well the system is performing in its current environment with respect to those requirements.

### Objective functions

We propose first a technique for specifying the requirements of a computer system, its subsystems, and its components to allow the efficient evaluation of the effects of the environment on the system. Note that we are not attempting to provide validation of the system, but only an efficient, quantifiable metric to measure changes in the environment and their effect on the system's ability to meet its requirements. This *objective function* is a partial specification of the system that uses derived (predictive) values or observed (reactive) values to represent the environment and computes a set of values for particular attributes that can be compared to a set of desired or expected values that represent the specification for those attributes.

The purpose of this function is to capture in a computable definition the services required of the system and the environmental factors under which those services can be expected to be delivered. The function must be able to express all objectives that affect its dependability, including *functionality*, the extent of the services provided by the system, *performance*, the time-related attributes, *security*, a measure of the system's resilience to maliciously (as opposed to inadvertently) introduced faults, *safety*, the requirements of the system that deal with the dangers that must be considered, and *precision*, which we can define as a measure of the latitude that will be accepted in the system's meeting its specified requirements.

There are two types of attributes that can be involved in such a function. *Factors* are attributes that are inherent to the system or component being described by the function and are not affected by changes in the environment. Those attributes that can vary in response to

changes in the environment are then the *parameters* to the function. Although in most cases the factors in an objective function will be fixed for a given component or system, we prefer the term "factor" to "constant" since a fixed value in one function may vary in a different function, and since the value of a factor may be assumed rather than known.

Examples of objective functions can be found in many forms, both practical and theoretical. A frequently encountered example of an objective function is the reliability measure. This kind of metric usually forms the basis for analysis of the predicted reliability of a system or component, based either on failure data observed in testing or on values predicted from modeling. Reliability measures can be combined with performance measures to describe an attribute sometimes called performability. Other metrics evaluate requirements in terms of availability, security, and precision.

Most often, dependability measures are used during the design and testing of a system, not its operation. Thus, the evaluation metrics are usually computed manually or by using automated tools. A few experimental operating systems provide implementations of a partial objective function that is used for scheduling and resource management based on the stated requirements and characteristics of the components of a system.

The objective function gives us a model for determining whether a system is behaving "acceptably" with regard to its requirements. We can compute values for the attributes from the function, using the parameters as a representation of the internal and external states of the system (i.e., of the current environment). If these computed values match the required values for the attributes, then the system meets its objectives under the current environment.

Furthermore, this model allows us to define violations of the requirements and gives us a method of detecting when they occur. If the computed value of an attribute does not match

its specified value, then we say that an *anomaly* in that attribute has occurred, indicating that the requirement for that attribute has not been met. Anomalies can be defined independently for all attributes specified in an objective function.

Notice that many types of errors can be detected using this model. Performance anomalies, such as missed deadlines or timeouts, are simply a case of a performance attribute (time) whose value exceeds a specified threshold. Functionality anomalies, like a data item that is not available, can be expressed as a value of zero for an attribute where a nonzero value is required. The advantage of our model is that we can consider a range of values for any attribute in an objective function and can thus specify a range of objective values that allow for a tolerance in meeting the requirement. This tolerance supports adaptivity by associating different strategies for fault management with different (possibly overlapping) ranges of attribute values. A variety of policies for managing adaptivity in a system can be described with this model as well, such as maximizing the value of the objective function or more complex policies such as minimizing the distance of all attribute values from their specified values or maximizing the likelihood that all attributes will meet their specified objectives.

### Consistency measures

A second consideration in specifying adaptivity in a survivable distributed system is the coordination of resources involved in the same computation. As mentioned earlier, one of the reasons that distributed computing is so important to the development of survivable systems is that it provides a basis for replication of the processing, data, and communication resources on which providing continued service in the presence of failures depends. However, the presence of multiple versions or copies of a resource implies that the requirements for that resource

apply to all versions. As we shall see, however, an adaptive system may allow replicated versions of a resource to have different *views* of the resource. The question then becomes how to specify the degree of similarity in the individual views.

To measure the extent of agreement or "sameness" of a view of a resource by a set of computations, we can apply a *consistency measure* to each common attribute of the objective functions among the computations. This measure, which we define qualitatively, provides for agreement among the views of a resource at three levels. *Nonmutual consistency* describes the situation in which a set of computations have completely independent views of a resource, although they are all constrained by the requirements pertaining to it. An example of this is concurrency control on a shared database, in which any computation may access the database independently, although each is constrained to leave it in a state that is consistent with some specified invariant.

*Mutual consistency* is a stronger condition in that there is implicit in its definition both replication of the resource and some relationship among the computations' views of the resource. In *mutual replicated consistency*, all views of a resource by a set of computations are identical and consistent with each other at all times. In *mutual nonreplicated consistency*, the views are not necessarily identical, but are consistent with respect to each other. For a database, mutual consistency implies that all computations accessing the database agree on what is in the database at any given moment. Mutual replicated consistency requires that the views be exactly the same as well, a condition that can be achieved using serialized transactions. Mutual nonreplicated consistency requires only that the computations agree on the contents of the database without necessarily seeing the complete database, as occurs in a relational database among multiple database servers responsible for maintaining each relation.

We can give examples of these levels of consistency for the processing resource class as well. Many fault management schemes rely on a replicated group of processes that cooperate to provide a single service. For this service to be continuously available, the process group must have the same view of the group membership at the same time. When the leader of a group fails, the others must know who belongs to the group at that time in order to agree on a new leader. This is therefore an example of mutually replicated consistency in processing.

Mutual nonreplicated consistency in processes is seen in the idea of a group of distinct processes communicating among themselves. We may choose to use process rollback and restart in the case of a process failure. In this situation, however, it is possible that the rollback of one process will trigger the rollback of a process that communicated with the first one in order to undo the effects of the communication between them. This second rollback might trigger the rollback of a third process, and so on, until the entire group has been rolled back to some point far beyond the point of failure of the original process. Although proposed techniques for limiting this "domino effect" vary, it is not necessary in any of them that every process know about every other process in the *conversation* in order to avoid the effect. Since the processes are related to the conversation, but not every process sees the entire group, such a scheme maintains consistency among the processes at the mutual, but nonreplicated, level.

Nonmutual consistency in processes is found in situations similar to that of the data example, only with resources or devices rather than data items. In most operating systems, scheduling of processes for execution on either a CPU or a group of processors is constrained by the requirement that processes give up a processor when waiting for a system function, such as I/O, to be completed. The processes are required to leave the processor in an orderly

state so another process can be executed.

Examples of the varying levels of consistency are also seen in communication. The idea of an atomic broadcast, in which all processes involved must see the same message or else none of them may see it, is constrained to be consistent at the mutual replicated level. A multimedia communication such as simultaneous voice and data transmission is an example of mutual nonreplicated consistency, in which the receiving processes can compare their views to synchronize the transmissions and assure correct reception. Reliable message passing over a network is a simple example of a nonmutually consistent requirement, in that all communication over the network is independent but must follow a transmission protocol to avoid interfering with each other.

## **CHOOSING A STRATEGY BASED ON REQUIREMENTS**

By combining a measure of consistency like that proposed above with the notion of the objective function, we can evaluate how well a group of computations in a distributed system is able to maintain the service(s) they are expected to provide. Specifying the objective values that must be met and the level of consistency that must be maintained specifies the bounds for both individual and group behavior that can be considered "acceptable." (Note that a single process providing a service is a degenerate case of nonmutual consistency.)

Of course, anomalies in the achievement of an objective, such as functionality, performance, or precision, can affect the system's ability to achieve the required level of consistency, and vice versa. Fault management strategies can prevent these anomalies either from occurring or from affecting the ability of the system to meet its other requirements. What we would like to do now is to determine which strategies can be applied under different environments to various anomaly classes so that we can maintain acceptable behavior in the system.

We would also like to be able to determine when this is not possible, so that we may consider changing strategies when conditions indicate it may be advantageous to do so.

Since an anomaly can be either detected or predicted, we may use them to either recover from or avoid errors and potential failures. The number and severity of the anomalies can guide the choice of a strategy. More importantly, however, the presence of anomalies indicates the ways in which the system is not behaving acceptably. It may be that the requirements for this behavior can be met if other requirements are relaxed or ignored. In other words, we may choose to change strategies based on tradeoffs among the system requirements, deemphasizing some to maintain others.

The idea of trading off one requirement to maintain another is not new. The most typical tradeoff is functionality against performance. Some systems maintain full functionality of tasks when processing or communication anomalies occur, either by providing full redundancy of all the tasks at all nodes or by dynamically modifying the scheduling to continue the task processing at another, reachable node. This kind of scheme tends to reduce the performance capabilities of a system for a given number of processors, especially in the second case if the processors were already being highly utilized. Other systems reduce functionality by dropping less "important" tasks as the ability to meet performance goals is reduced. This scheme is typically found in real-time systems, where certain tasks identified as critical must have every attempt made to schedule them by their deadline, even if other tasks must be delayed or deleted.

In many types of dependable systems, such tradeoffs between objective requirements are sufficient to provide the required level of service under a narrowly defined, highly predictable range of adversity. In survivable systems, however, it is necessary to consider the possibility



of widespread, clustered failures that cannot be compensated for by such methods. If giving up objectives is insufficient to allow a system to survive such loss, then we must consider giving up consistency instead. This approach can be found in distributed systems that handle partitioning of the system by supporting independent computations that may later be reconciled to restore consistency among them.

We therefore present another categorization of fault management strategies based on this notion of relaxing or trading off requirements in exchange for continued survival of the system. We define three policies under which strategies may be applied: *optimistic*, for environments in which so few anomalies are expected that they can be explicitly detected and handled without long-term effects on other requirements of the system; *pessimistic*, for environments in which a sufficient number or frequency of anomalies are expected as to require the relaxing of one or more of the objective requirements of the system to avoid their causing other anomalies and possibly catastrophic failure; and *ultrapessimistic*, for environments in which the anomalies that have occurred require the relaxation or abandonment of consistency requirements in order to preserve the critical services of the system.

Optimistic strategies depend on explicit detection of and recovery from anomalies in the system. Their usual response to detection of an anomaly is restoration of a consistent system state, followed either by an attempt to repeat the operation that failed or by reconfiguration of the system to include an alternative version or spare. This is usually done in a single-threaded mode, so that the recovery can take place with the minimal allocation of resources and with the minimal disruption of other computations. Since consistency among a group of related computations must be preserved, often these approaches require a protocol for rollback of multiple processes.

Clearly, this kind of strategy is sufficient when error rates are low and there are sufficient resources (including time) to recover from anomalies. However, as error rates increase, or the environment becomes such that recovery would affect the operation of the system too greatly, optimistic strategies become ineffective or too costly. Instead, we need to consider using pessimistic strategies that avoid anomalies by relaxing one or more of the objective requirements in order to maintain the other requirements, including consistency. Typically, these strategies operate in a multi-thread fashion and avoid failures by scheduling or masking of anomalies. This kind of approach will handle a higher error rate, but tends to use resources. By relaxing the objective requirements, we can continue to operate and avoid failures, although we pay the price in loss of functionality, performance, or precision.

We may also choose to give up consistency to maintain the specified objectives. This is not a widely used approach, since the notion of consistency is seldom seen as a requirement that can be specified at different levels. In this situation, anomalies can be masked or even ignored with respect to their effect on consistency, so recovery is kept to a minimum. Degradation of service may be required, but we can determine which services to eliminate, if we know which ones are related to maintaining the objectives and which to consistency. It would appear that this kind of approach would tolerate fairly high error rates, since it allows the system to ignore certain kinds of errors and requires little resource overhead. In fact, the relaxation of consistency requirements, coupled with the degradation of service, may provide enough processing power for the system to recover some of its objectives (particularly functionality) on its own, without explicit recovery mechanisms.

One reason that ultrapessimistic approaches have not been studied more may be that they are hard to describe in non-application specific terms. It is difficult to determine the most

critical functions of an application and where consistency may be safely given up without a significant amount of *a priori* information about the application. Less information of this sort is needed for pessimistic strategies, since avoidance techniques can often make use of modeling and prediction to determine the parameters of the objective function. Optimistic approaches require the least *a priori* information of all, since the application can use information collected from execution to estimate rather than predict the parameter values.

The identification of these three categories for fault management strategies allows us to consider which strategies may be most effective when the surrounding environment rapidly and drastically changes. For example, a radar system may operate in three modes: ready mode, in which few errors are expected and can be repaired at relative leisure; alert mode, in which errors must be avoided due to the need to maintain the system in full operation; and conflict mode, in which the system is experiencing high losses of components and has limited resource and replacement capabilities. We could choose to use an optimistic strategy in the first situation, a pessimistic strategy in the second, and an ultrapessimistic strategy in the third.

## A TAXONOMY FOR ADAPTIVE FAULT MANAGEMENT

We can summarize this discussion as a taxonomy for fault management strategies. Previous taxonomies related to this problem have not considered adaptivity as an issue, concentrating instead on comparing specific hardware or software architectures, or else they provide a high-level view of dependable systems that is of little use in the detailed design and analysis of such systems.

A taxonomy that considers adaptivity is shown in Figure 2. This taxonomy allows us to classify strategies for fault management according to the resource classes to which they apply and the level of consistency that they maintain. Within these categories we can define both

the anomalies that can indicate that an objective requirement has been violated, so we can detect the need for a new strategy, and the strategies that may be applied for each policy.

#### RESOURCE CLASS (Processing, Data, Communication)

CONSISTENCY	ERROR DETECTION	STRATEGIES
Mutual replicated	Functional anomalies	Optimistic
	Performance anomalies	Pessimistic
	Precision anomalies	Ultrapessimistic
Mutual nonreplicated	Functional anomalies	Optimistic
	Performance anomalies	Pessimistic
	Precision anomalies	Ultrapessimistic
Nonmutual	Functional anomalies	Optimistic
	Performance anomalies	Pessimistic
	Precision anomalies	Ultrapessimistic

Figure 2. A taxonomy for adaptive fault management strategies.

The advantage of such a taxonomy is that it can be applied to a particular problem to define the strategy classes that are best for a given set of circumstances. A design paradigm for survivable systems can be envisioned that uses these classifications, first, to guide how the system can be constructed to support adaptation among the strategies and, second, to define the classes of strategies that best fit all the environments that the system is expected to encounter (even if the likelihood of encountering any of them is small).

The proposed taxonomy encompasses many of the well-known fault management strategies for both software and hardware. However, most of these strategies fall into the optimistic or pessimistic categories. Very few strategies exist that utilize the relaxation of consistency requirements to maintain the objectives. In survivable systems that are deployed in potentially harsh environments, such strategies may prove to be extremely important.

## Conclusion

We have presented a number of categories for fault management strategies that can be used to guide their selection in an adaptive survivable system. By using the guidelines, the system designer will be able to define a number of situations under which different strategies may be effectively used, thus improving the chances of the system avoiding catastrophic failure brought on by the violation of its assumptions.

Clearly, much research is needed in this area to make adaptivity a tool in the design and operation of survivable systems. First and foremost, the taxonomy must be populated with the known strategies to define how well we understand the various categories. It appears, based on preliminary work, that we have many techniques for optimistic and pessimistic fault management in the areas of data and communication. Processing fault management has made much progress in the area of optimistic strategies, but less so in the area of pessimistic. None of these areas have been extensively studied with respect to ultrapessimistic strategies.

Practical issues also need to be considered. The major problem with the use of objective functions is their definition in a complete, yet operationally tractable manner. We need to be able to define objective functions for the requirements (at least those related to fault management) in such a way that they can be evaluated. We must also consider how and when to make the transition between strategies, which will involve evaluating the tradeoffs between objective and consistency requirements. It is not yet clear if this can be done other than on an application-specific basis, although general principles should be definable. Similar problems exist for the consistency measures.

Other considerations include how to choose the appropriate strategy from among those available. We will need to find an efficient rule base and evaluation method for deciding

among strategies in every system designed in this way. The rules for making such decisions will be complex and a general, or at least automatable, decider will be complex as well. Furthermore, to avoid a single failure point in the decision manager, it will be necessary to reliably distribute this function over the nodes. While a number of operating systems claim to support reliable distributed computing, the evaluation of their support for survivable systems based on the approaches described here will require significant effort.

We have mentioned that it may be possible to base a design paradigm for adaptive survivable C3 systems on this taxonomy. It appears, although we have not yet verified, that the approaches here can be used for subsystem as well as system design and implementation. The primary questions are how to integrate subsystems back into the larger system and how to control adaptivity among multiple concurrent systems in the same application. The relationships among processing, data, and communication may not be orthogonal, and the specification and management of adaptivity at the lower levels will need to be considered carefully to allow the interaction among the subsystems to be managed.

Eventually, we hope that the work outlined here will provide a basis for survivable systems that can be "engineered" rather than created. We fully agree that "... building distributed fault-tolerant systems will remain an art in the foreseeable future" [Cristian91], but we are confident that study such as this will help in defining the skills needed to achieve more highly dependable computing systems at a time not as distant as we might otherwise believe.

## ACKNOWLEDGEMENTS

The author wishes to thank Thomas F. Lawrence of Rome Laboratory for the ideas on which this report was based and for the helpful discussions on the taxonomy.

This work was supported by the AFOSR Summer Faculty Research Program. The views and opinions contained in this report are those of the author and should not be construed as an official Department of Defense position, policy, or decision.

## REFERENCES

- [Avizienis89] A. Avizienis (ed.), *Application of Fault Tolerance Technology: Design of Fault-Tolerant Systems*, BM/C3 Algorithm and Processor Working Group Report, Rome Air Development Center, Griffiss AFB, New York (October 1989).
- [Cristian91] F. Cristian, Understanding Fault-Tolerant Distributed Systems, *Comm. ACM* 34,2 (February 1991), 56-78.
- [Kohler81] W.H. Kohler, A Survey of Techniques for Synchronization and Recovery in Decentralized Computer Systems, *Computing Surveys* 13,2 (June 1981), 149-183.

## APPENDIX 1

### AN ANNOTATED BIBLIOGRAPHY OF DEPENDABLE DISTRIBUTED COMPUTING

Rex E. Gantenbein, Associate Professor  
Department of Computer Science  
University of Wyoming  
P.O. Box 3682  
Laramie WY 82071

#### 1. OVERVIEW OF DEPENDABILITY ISSUES

##### Textbooks

T. Anderson (ed.), *Resilient Computing Systems, Volume I*, John Wiley and Sons (1985).

An excellent overview of the major issues of dependability. The chapters, each written by an expert in his or her field, cover hardware, software, and communication reliability, real-time and distributed systems, safety, reliability measures, and case studies of commercial systems circa 1985. I don't know what ever became of Volume II.

B.W. Johnson, *Design and Analysis of Fault-Tolerant Digital Systems*, Addison-Wesley (1989).

An introduction to the design and analysis of dependable hardware. The emphasis is on techniques and concepts, and the author claims that "no previous knowledge" of fault tolerance is needed. One chapter deals with VLSI, and several sample designs are included as examples in the text.

B. Littlewood (ed.), *Software Reliability: Achievement and Assessment*, Blackwell Scientific Publications (1987).

Another collection of summaries written by a collection of experts, with the emphasis on software. Much of the book deals with reliability measurement, but other issues such as fault tolerance and software safety are addressed in chapters as well.

V.P. Nelson and B.D. Carroll (eds.), *Tutorial: Fault-Tolerant Computing*, IEEE Comp. Soc. Press (1987).

One of the IEEE tutorials. This is a collection of papers on a wide range of fault-tolerance subjects. Although slightly dated now (the papers all date back to 1986 or earlier), this tutorial is still a thorough overview of the pioneering work in many areas of dependability.

D.K. Pradhan (ed.), *Fault-Tolerant Computing: Theory and Techniques, Volume I and II*

A two-volume set. Volume I deals primarily with circuit level techniques, including test generation, designing for testability, fault simulation, and coding theory and techniques. Volume



II deals with system-level techniques, including architecture (general and multiprocessor), diagnosis, reliability estimation tools and techniques, and software fault tolerance. I prefer the Anderson book as an introduction, but this is a good reference for more esoteric issues and covers the circuit-level issues in detail.

### Articles

A. Avizienis and J.-C. Laprie, Dependable Computing: From Concepts to Design Diversity, *Proc. of the IEEE* 74,5 (May 1986), 629-638.

An invited paper in a special issue on fault tolerance in VLSI. The paper defines the basic terms of fault tolerance and provides a preliminary, high-level taxonomy of dependability. Several approaches to the problem of fault tolerance are considered, with the emphasis placed, as would be expected from Avizienis, on *design diversity* as a means of tolerating design faults.

F. Cristian, Understanding Fault-Tolerant Distributed Systems, *Comm. ACM* 34,2 (February 1991), 56-78.

A comprehensive paper containing both concepts and case studies in fault-tolerant distributed computing. The paper presents a unified discussion of the problems to be solved at the hardware and software architecture levels as well as approaches to their solution based on both avoidance (masking) and tolerance (recovery). It is suggested in the paper that the key to dependability is balancing failure detection, recovery, and masking redundancy at various levels of system abstraction, so that the provision of dependability at lower levels will make the higher levels easier to design.

J. Gray, Why Do Computers Stop and What Can Be Done About It?, *Proc. 5th Symp. on Reliability in Dist. Software and Database Sys.* (July 1986), IEEE Comp. Soc. Press, 3-12.

A report on the results of studies on failures in Tandem systems, for whom the author worked at the time. This paper is often quoted as a source of data on reliability in commercial fault-tolerant systems, even though the data was collected from user problem reports and as such is, Gray admits, not very precise and probably underreported. Among the conclusions is that administration and software were the major causes of failure in these systems.

J.G. Kuhl and S.M. Reddy, Fault-Tolerance Considerations in Large, Multiple-Processor Systems, *Computer* 19,3 (March 1986), 56-67.

A survey of methods and techniques for achieving hardware fault tolerance in large, multi-computer systems, focusing on distributed control. The paper gives a framework for hardware fault tolerance in distributed systems and discusses replication and masking techniques, diagnosis, repair and recovery, and communication issues.

J.-C. Laprie et al., Definition and Analysis of Hardware- and Software-Fault-Tolerant Architectures, *Computer* 23,7 (July 1990), 39-51.

A review of software fault tolerance techniques, including recovery blocks, N-version programming, and hybrid schemes referred to here as *N self-checking programs*. Faults in software differ significantly from those in hardware, and architectural designs that attempt to

handle both are complex. Some highly abstract solutions to this problem are presented here, along with an evaluation of their effectiveness based on Markovian models of system behavior.

V.P. Nelson, Fault-Tolerant Computing: Fundamental Concepts, *Computer* 23,7 (July 1990), 19-25.

A concise overview of fault tolerance concepts, mechanisms, and strategies. The review is hardware-oriented, although many of the concepts can be applied to software as well. This paper, as well as the Laprie paper above, are part of a special issue of *Computer* on fault tolerance that covers a wide range of topics.

### Case studies of commercial systems

J.-P. Banatre et al., The Design and Building of Enchere, A Distributed Electronic Marketing System, *Comm. ACM* 29,1 (January 1986), 19-29.

A description of a French system that supports distributed synchronization, atomic activities, and a commit protocol with recovery algorithms.

P.A. Bernstein, Sequoia: A Fault-Tolerant Tightly Coupled Multiprocessor for Transaction Processing, *Computer* 21,2 (February 1988), 37-45.

A system with a multicomputer architecture that uses hardware fault detection and software recovery in the operating system.

D. Gifford and A. Spector, The Cirrus Banking Network, *Comm. ACM* 28,8 (August 1985), 798-807.

One of the series of case studies published by Gifford and Spector in the *Communications* during this period. This is actually an interview with the president of CIRRUS Systems, Inc., with discussion of the high availability features of the wide-area network that supports the transaction processing.

J.N. Gray and M. Anderson, Distributed Computer Systems: Four Case Studies, *Proc. of the IEEE* 75,5 (May 1987), 719-726.

A comparison of four distributed architectures with respect to fault management. The comparison centers around decentralization of hardware, control, and redundancy management in varying degrees.

E.S. Harrison and E.J. Schmitt, The Structure of System/88, A Fault-Tolerant Computer, *IBM Sys. Journal* 26,3 (March 1987), 293-318.

Describes the configurations of the System/88 family of IBM products, which are based on the Stratus/32 system. The system achieves fault tolerance via hardware duplexing coupled with a distributed operating system that replicates resources transparently across several nodes or a network of multinode systems.

D.P. Siewiorek, Fault Tolerance in Commercial Computers, *Computer* 23,7 (July 1990), 26-37.

An survey of fault tolerance in commercial systems as of 1990. The systems reviewed include uniprocessor systems (VAX 8600 and IBM 3090), multicomputer systems (Tandem, Stratus, and VAXft 3000), and multiprocessor systems (Teradata and Sequoia). The systems are evaluated as to the sources of the errors that they handle and the approaches used. This paper is found in the special issue of *Computer* mentioned previously.

J.J. Wallace and W.W. Barnes, Designing for Ultrahigh Availability: The UNIX RTR Operating System, *Computer* 17,8 (August 1984), 31-39.

A study of the UNIX RTR operating system solutions to the problem of providing extremely high availability. This system is used by AT&T in support of its ESS telephone switching systems, which are famous for their (expected) downtime on the order of a few minutes per year, including maintenance.

## 2. DEPENDABLE DISTRIBUTED RESOURCE MANAGEMENT

### Data

P.A. Bernstein and N. Goodman, Concurrency Control in Distributed Database Systems, *Computing Surveys* 13,2 (June 1981), 165-201.

A survey of 48 methods for concurrency control in distributed databases. The paper primarily considers mechanisms and techniques for correctly synchronizing concurrent access to data, with performance issues given only secondary consideration. Both centralized and distributed schemes are discussed. Of particular interest is their identification of "anomalies" in consistency in the data and the methods by which these can be eliminated (serialization, etc.).

K.P. Eswaran et al., The Notion of Consistency and Predicate Locks in a Data Base System, *Comm. ACM* 19,11 (November 1976), 624-633.

The paper that really started it all with respect to concurrency control on shared data. Is there a graduate student in computer science/engineering who hasn't read this? The paper presents the idea of consistency for databases and how it can be preserved using predicate locks. Although many other schemes have improved on these ideas, this is still worth the effort to read for background in the area.

H. Garcia-Molina and R.K. Abbott, Reliable Distributed Database Management, *Proc. of the IEEE* 75,5 (May 1987), 601-620.

Algorithms and techniques for achieving reliability in dispersed and/or replicated data. Some methods for handling partitioning of a database are also briefly surveyed.

M. Herlihy, Comparing How Atomicity Mechanisms Support Replication, *Proc. 4th ACM Symp. on Principles of Dist. Computing* (August 1985), ACM Press, 102-110.

Compares the constraints on replication necessary to maximize the concurrency in a distributed database system. The author suggests that availability of replicated data should be the criterion by which atomicity mechanisms are evaluated. The main result presented is that

hybrid schemes that utilize both locking and timestamps permit more concurrency than locking alone.

W.H. Kohler, A Survey of Techniques for Synchronization and Recovery in Decentralized Computer Systems, *Computing Surveys* 13,2 (June 1981), 149-185.

A comprehensive survey of techniques for synchronizing access to shared objects in a distributed system and recovering those objects after failures. The emphasis is on software structuring techniques to achieve error recovery. Although the paper is somewhat dated, these techniques have been extensively used.

J.D. Noe and A. Andreassian, Effectiveness of Replication in Distributed Computer Networks, *Proc. 7th Int. Conf. on Dist. Computing Sys.* (September 1987), IEEE Comp. Soc. Press, 508-513.

A comparison of voting, regeneration, and available copies schemes for replication control. The schemes are evaluated on the basis of storage costs for replicated copies. The result presented is that all these methods increase the availability of the data, but little gain is seen with more than two copies, making voting (which requires at least three copies) less preferable than the other schemes.

F. Pitelli and H. Garcia-Molina, Recovery in a Triple Modular Redundant Database System, *Proc. 7th Int. Conf. on Dist. Computing Sys.* (September 1987), IEEE Comp. Soc. Press, 514-520.

A description of a method for providing a replicated database that can survive failures at one of its nodes. The system will continue processing transactions at the other nodes, while the lost node upon repair will take a snapshot of the corrupted parts of the database and initiate a "catchup" process to get back into synchronization with the others. The motivation for this approach comes from results suggesting that pure recovery in this system hinders its ability to process transactions efficiently.

R. Strong et al., Handshake Protocols, *Proc. 7th Int. Conf. on Dist. Computing Sys.* (September 1987), IEEE Comp. Soc. Press, 521-528.

Presents a paradigm and techniques for replicated data management called *synchronous distributed memory*. The emphasis is on maintaining consistency in the data by detecting and recovering from failures rather than avoiding it, especially in the case of transient partitions in a network. A useful set of protocols is described that provide ways to reestablish consistency in the data once it has been lost.

J.S.M. Verhofstad, Recovery Techniques for Database Systems, *Computing Surveys* 10,2 (June 1978), 167-195.

A old, but comprehensive for the time, survey of methods for maintaining consistency and availability in databases. Since most of these methods are still in use, this article is still valuable despite its age.

Communication

K. Birman and T. Joseph, *Reliable Communication in the Presence of Failures*, *ACM Trans. on Computer Sys.* 5,1 (February 1987).

Describes techniques for the design and verification of a distributed communication facility. The methods support a family of reliable multicast protocols that respect a variety of ordering constraints and thus have varying consistency and performance depending on the degree of ordering that must be maintained. The protocols assure that the processes in a group will observe consistent orderings of events affecting the group, where consistency does not necessarily imply identity or simultaneity. This paper, and the others by Birman, describe techniques used in the ISIS distributed operating environment developed at Cornell.

K.P. Birman and T.A. Joseph, *Exploiting Replication*, Tech. Report TR 88-917, Cornell Univ. Dept. of Computer Science (June 1988).

A description of communication among replicated processes in the ISIS distributed system. The mechanisms provided in ISIS allow varying levels of "synchrony" among processes that range from lockstep to *virtual synchrony*, a level of consistency in which communication is considered synchronized with respect to message ordering as long as the results are indistinguishable in the receiving processes.

### Processing

K.P. Birman et al., *Implementing Fault-Tolerant Distributed Objects*, *Proc. 4th Symp. on Reliability in Dist. Software and Database Sys.* (October 1984). IEEE Comp. Soc. Press.

A reference for replication control in ISIS. The system uses an available copies scheme that removes failed nodes from the group and reconfigures them back in once they have recovered. The scheme uses the communication protocols described in the other papers listed above to support varying levels of consistency among the group membership.

B.A. Coan and G. Thomas, *Agreeing on a Leader in Real-Time*, *Proc. Real-Time Sys. Symp.*, (December 1990), IEEE Comp. Soc. Press, 166-172.

Protocols for selecting the leader of a "replication ring" of processes in real time. The method is intended for shared memory systems where a group of processes elect a leader that provides a service to requestors, although the service may be handled by all members of the ring. If the leader fails, a new leader can be elected in a bounded time using the methods described here.

F. Cristian, *Agreeing on Who is Present and Who is Absent in a Synchronous Distributed System*, *Proc. 18th Int. Symp. on Fault-Tolerant Computing* (June 1988), IEEE Comp. Soc. Press, 206-211.

Methods for detecting departures and joins in a process group and reliably reaching agreement on membership in a bounded time. The paper presents two protocols for this problem, one of which has fast detection of changes but high message traffic overhead, even when no changes occur. The second protocol has a longer delay but has a provable limit on message overhead in the absence of failures.

F. Cristian, A Probabilistic Approach to Distributed Clock Synchronization, *Proc. Real-Time Sys. Symp.* (1989), IEEE Comp. Soc. Press, 288-296.

A description of an approach to clock synchronization that defines the probability of getting distributed clocks to agree within a defined precision. This paper is interesting because it exemplifies the issues of tradeoffs for both objectives and consistency in an application. The protocols presented can tolerate processor and communication failure and detect clock failures, but the tradeoffs between message traffic, time to synchronize, and the probability of successful synchronization within a specified time are such that the same strategy may have very different results depending on the environment.

M.J. Fischer and N.A. Lynch, A Lower Bound for the Time to Assure Interactive Consistency, *Information Processing Letters* 14,4 (June 1982), 183-186.

A proof of the property that for  $n$  processors, of which at most  $m$  are faulty,  $m+1$  rounds of communication are required for interactive consistency, as defined in the paper by Pease et al. below. These are two of three papers listed here that discuss the issues of consistency among processes in the context of the Byzantine generals problem (the third, by Lamport et al., actually defines the problem).

E.D. Jensen, C.D. Locke, and H. Tokuda, A Time-Driven Scheduling Model for Real-Time Operating Systems, *Proc. Real-Time Sys. Symp.* (December 1985), IEEE Comp. Soc. Press, 112-122.

Describes a time-driven model for process scheduling that includes priority as a function of time. This *time-value function* is supported in the Alpha distributed operating system and, as such, is one example of an implementable objective function. The paper presents the results of experiments in evaluating the effectiveness of several scheduling algorithms using this model, including the *best-effort scheduler* that is described in Locke's thesis below and implemented in the first releases of Alpha.

M.K. Joseph and A. Avizienis, A Fault Tolerance Approach to Computer Viruses, *IEEE Symp. on Security and Privacy* (April 1988), IEEE Comp. Sci. Press, 52-58.

Examples of using fault tolerance as an approach for security violations. This is an largely unexplored area, although it seems feasible if one views security as the resilience to anomalies in the system due to maliciously (rather than inadvertently) introduced faults. Since the anomalies may result in denial of services or corruption of sensitive information, security requirements can in principle be specified in the same way as other requirements such as functionality and performance.

M.K. Joseph, *Integrating Security into Current Dependability Concepts and Models*, LAAS-CNRS Tech. Report 89194, Toulouse, France (June 1989).

Qualitative models for approaches that combine dependability and security as outlined in the reference for Joseph and Avizienis above.

L. Lamport, R. Shostak, and M. Pease, The Byzantine Generals Problem, *ACM. Trans. on Programming Lang. and Sys.* 4,3 (July 1982), 382-401.

Description of the "Byzantine generals" problem, which is essentially the problem of reaching agreement among a group of processors in the presence of faults in its members. The paper proves that, if oral (that is, unreliable) messages are used for communication, withstanding  $m$  failures requires a minimum of  $3m+1$  processors, then presents a solution to the problem that can be implemented with the minimum number of processors. A second solution that uses written and signed messages (analogous to authenticated or tagged messages) allows a reduction in the number of processors needed to withstand a given number of faulty processors. The Byzantine generals problem is an example of maintaining mutual replicated consistency among a group of processors.

C.D. Locke, *Best-Effort Decision Making for Real-Time Scheduling*, Ph.D. Thesis, Department of Computer Science, Carnegie-Mellon University (May 1986).

A description of scheduling algorithms in real-time systems, including the best-effort scheduler used in Alpha. This provides an example of how an objective function (in this case, the time-value function) can be used in a system to evaluate the ability of the system to meet its requirements (in this case, the real-time deadlines). When the system load increases to the point that some degradation is required, the best-effort scheduler uses the time-value functions for all tasks to maximize the "value" of the system with respect to the task schedule chosen, rather than scheduling strictly on priority or criticality as is typically done.

S.R. Mahaney and F.B. Schneider, Inexact Agreement: Accuracy, Precision, and Graceful Degradation, *Proc. 4th Ann. ACM Symp. on Principles of Dist. Computing* (August 1985), ACM Press, 237-249.

Two protocols that allow processes to "agree" on a value without requiring that the value be the same in both. The protocols let the processes exchange their values and use them to compute new ones that should be closer to both the actual value and to each other's. As long as fewer than one-third of the processes are faulty, then these protocols will result in convergence for the value (if not, then the divergence is bounded). This is yet another example of levels of consistency.

R. Obermarck, Distributed Deadlock Detection Algorithm, *ACM Trans. on Database Sys.* 7,2 (June 1982), 187-208.

An algorithm for detecting deadlock in distributed transactions. Centralized algorithms require a site to be in contact with all other users, while many distributed algorithms replicate the "wait-for" graph used to detect deadlock at all sites. This scheme splits the detection graph among the sites, providing different views of the processing at each. Local deadlocks can be broken locally, and only global deadlocks involving transactions at remote sites need to be coordinated. This algorithm is susceptible to false alarms, so it may abort some transactions needlessly. An example of mutual non-replicated consistency.

M. Pease, R. Shostak, and L. Lamport, Reaching Agreement in the Presence of Faults, *Journal of the ACM* 27,2 (April 1980), 228-234.

The original proof that algorithms for interactive consistency among a group of  $n$  processors with  $m$  or fewer faulty processors can only be devised if  $n$  is  $3m+1$  or greater. The definition of interactive consistency given here is that all nonfaulty processors must come to a consistent

view of the values held by all processors, including the faulty ones. Note that the faulty processors are not constrained by this requirement, which is then weaker than a strict (and unattainable, in all likelihood) requirement for consistency in all processors, faulty or not.

### 3. DESIGNING DEPENDABLE DISTRIBUTED SYSTEMS

#### Design paradigms

A. Avizienis (ed.), *Application of Fault Tolerance Technology: Design of Fault-Tolerant Systems*, BM/C3 Algorithm and Processor Working Group Report, Rome Air Development Center, Griffiss AFB, New York (October 1989).

A report prepared for RADC based on the work performed by a group of experts in 1986-1987. The report discusses many of the important issues in the application of fault tolerance to distributed systems. The most interesting discussion centers around a design paradigm for the development of dependable systems, the essence of which is the partitioning of a system design into subsystems for which the error detection and fault management techniques are individually devised and evaluated. While it is hard to call this a radical notion, the idea of a design paradigm that supports a structured method for the development of dependable systems is important. From a software engineering point of view, it is impossible to develop truly dependable systems unless the development process itself is dependable and evaluable.

F.G.F. Davis and R.E. Gantenbein, Responding to Catastrophic Errors: A Design Technique for Fault-Tolerant Software, *Journal of Sys. and Software* (to appear 1991).

A design paradigm for systems that may experience catastrophic errors. The intent is to identify those failures that cannot be handled in a localized manner and then design the system to avoid their occurrence.

I. Dunham, *Abstraction and Methodical Development of Fault-Tolerant Software*, Tech. Report CMU-CS-86-112, Carnegie-Mellon Univ. (1986).

A Ph.D. thesis from CMU describing a structured approach to fault-tolerant software design. This is one of the few documented attempts to use software engineering-like techniques for dependable software. Unfortunately, the only place I've seen it mentioned is in an IBM University-Level Computer Science course on fault-tolerant software that I teach.

N.G. Leveson, Building Safe Software, in *Software Reliability: Achievement and Assessment* (B. Littlewood, ed.), Blackwell Scientific Publishing, Oxford, 1987.

An approach to designing systems with the emphasis of preserving safety as an attribute as well as the other objectives normally considered. In safety-critical systems, operational failure can result in damage to property, personnel, or the environment. The approach described here uses risk factors to determine where the need for dependability (either in fault avoidance or tolerance) is greatest. While intended more for control or avionics applications than survivable systems, this approach has merit in its emphasis on effective application of fault management as a means to prevent undesirable behavior instead of a means to preserve the *status quo*.

#### Hardware designs



R.E. Harper, J.H. Lala, and J.J. Deyst, Fault Tolerant Parallel Processor Architecture Overview, *Proc. 18th Int. Symp. on Fault-Tolerant Computing*, (June 1988), IEEE Comp. Soc. Press, 252-257.

An early description of the FTPP from Draper Labs. In many systems, reliability and throughput are seen as conflicting requirements, so this system was designed using clusters of multiprocessors. Of special interest here is the recognition that mutual synchrony is difficult to achieve with any efficiency, so they support nonhomogenous *functional* synchrony to coordinate equivalent execution among the clusters.

J.H. Lala and L.S. Alger, Hardware and Software Fault Tolerance: A Unified Architecture Approach, *Proc. 18th Int. Symp. on Fault-Tolerant Computing*, (June 1988), IEEE Comp. Soc. Press, 240-245.

A design based on clusters of multiprocessors that tolerates arbitrary hardware faults and efficiently implements N-version programming. This design adds an application processor to each multiprocessor cluster so that hardware and software faults can be differentiated and isolated. Masking is carried out by a configuration voter that carries the error history of each version, so that after a while it "learns" which versions are trustworthy and will decide on that basis rather than on a strict plurality.

J.H. Lala, R.E. Harper, and L.S. Alger, A Design Approach for Ultrareliable Real-Time Systems, *Computer* 24,5 (May 1991), 12-22.

A description of approaches to masking errors and achieving congruency among redundant hardware components. The context of the discussion is the design of architectures that can tolerate a single, arbitrary hardware fault as well as some double simultaneous faults without complete processor redundancy. (Note: I admit to being little concerned with strictly hardware-based techniques for fault tolerance, which is why this section is so sparsely populated. The basic issues of hardware fault tolerance are described in some of the overview papers, and management of multiple processors is included under processing in the previous section. However, this paper is a demonstration that fail-stop processors do indeed exist, making higher-level techniques feasible.)

### Software designs

T.E. Bihari and K. Schwan, A Comparison of Four Adaptation Algorithms for Increasing the Reliability of Real-Time Software, *Proc. Real-Time Sys. Symp.* (December 1988), IEEE Comp. Soc. Press, 232-241.

A description of an adaptive fault-tolerance manager in a real-time system. The manager incorporates a uniform model of fault-tolerance techniques and provides tools to implement algorithms for choosing and performing adaptation in real time. The adaptations include changes in the software configuration, its internal structure and function, or the software-to-hardware mapping. The system schedules fault tolerance (which may consist of redundant software, retry, or assignment) to favor those components that have the highest value of priority times number of failures (this is the objective function). Various combinations of fault-tolerance algorithms were tried, and the paper summarizes the average service value lost per fault in each case. Although the number of adaptations possible is small, this is an example of what adaptive fault management can achieve.

A.K. Caglayan, P.R. Lorzak, and D.E. Eckhardt, An Experimental Evaluation of Software Diversity in a Fault-Tolerant Avionics Application, *Proc. 7th Symp. on Reliable Dist. Sys.* (October 1988), IEEE Comp. Soc. Press, 63-70.

Algorithm diversity as applied to redundancy management software for a fault-tolerant sensor array. The study shows the performance gains achieved from using three diverse algorithms in parallel for error detection and isolation. The gains were mostly due to the elimination of false alarms. The paper contains a useful discussion of "median" versus "majority" voting to detect and mask errors.

W.R. Dunn and L.D. Corliss, Software Safety: A User's Practical Perspective, *Proc. Ann. Reliability and Maintainability Symp.* (January 1990), IEEE, 430-435.

An evaluation of software projects at the NASA Ames Research Center. Although the projects described here have exhibited high reliability in practice, the authors conclude that the design processes for dependable software must also be improved if any kind of guarantee is to be given. The examples demonstrate how ultrareliability was achieved by fault avoidance through design adaptation (in other words, they ran the system to see how it failed and then adapted the design or the implementation to prevent that failure from occurring again).

O. Gudmundsson et al., MARUTI: A Hard Real-Time Operating System, *Proc. 2nd IEEE Workshop on Experimental Dist. Sys.* (October 1990), IEEE Comp. Soc. Press, 29-34.

An overview of a distributed operating system designed at Maryland to reliably support hard real-time applications. The system provides fault-tolerant operation through mechanisms that can be used to implement a number of policies. The system, which has been prototyped on top of UNIX, provides for the definition of objects whose description in the system contains a specification of the fault tolerance technique used. It isn't clear from the paper, however, whether the technique used for an object is dynamic or static.

M.D. Hansen, Survey of Available Software-Safety Analysis Techniques, *Proc. Ann. Reliability and Maintainability Symp.* (January 1989), IEEE, 46-49.

A survey of techniques for conducting software safety analysis. These techniques are obviously adapted from NSCCA hardware analysis techniques (what's the point of calling a software design review "sneak circuit analysis?"), but make the point that structured reviews of code can eliminate a number of faults before they become operational. The paper suggests that Petri net modeling and simulation may be promising techniques for analyzing (and perhaps evaluating) a number of objectives, including performance, correctness, fault tolerance, and safety.

E.D. Jensen and J.D. Northcutt, Alpha: A Non-proprietary OS for Large, Complex, Distributed Real-Time Systems, *Proc. 2nd IEEE Workshop on Experimental Dist. Sys.* (October 1990), IEEE Comp. Soc. Press, 35-41.

A concise and easily accessible reference for the Alpha operating system developed at Carnegie-Mellon. Alpha has many properties that support the development of survivable distributed systems, including the provision of time-value functions which, as mentioned in the previous section, are an example of an implementable objective function for performance, functionality, and precision with respect to scheduling.

B. Littlewood and D.R. Miller, A Conceptual Model of Multi-Version Software, *Proc. 17th Int. Symp. on Fault-Tolerant Computing* (June 1987), IEEE Comp. Soc. Press, 150-155.

A study of how diversity of development methods affects the resilience of multi-version software systems. The value of diverse development is shown to depend on the characteristics of the system, particularly in the effects of similar errors.

T.P. Ng, The Design and Implementation of a Reliable Distributed Operating System - ROSE, *Proc. 9th IEEE Symp. on Reliable Dist. Sys.* (1989), IEEE Comp. Soc. Press, 2-11.

A brief description of an experimental distributed operating system being developed at Illinois. The system, which is based on the V kernel with extensions, provides tasks that can detect when other tasks have become unreachable and an abstraction of replicated tasks that can be utilized by applications. Of particular interest is the *guarantee* function, which guarantees that a task will be in a state consistent with having sent a message (unless it has been terminated due to a sufficient number of failures in sending the message) and thus supports mutual consistency among the tasks.

K. Schwan, A. Gheith, and H. Zhou, From CHAOSbase to CHAOSarc: A Family of Real-Time Kernels, *Proc. Real-Time Sys. Symp.* (December 1990), IEEE Comp. Soc. Press, 82-91.

Presents a family of object-based operating system kernels that are extensible and customizable to a variety of real-time environments. The major feature of this family of CHAOS kernels is the support of object abstractions that can be adaptively specified with respect to a number of attributes. The external interface to any CHAOS object is a policy abstraction that interprets the current specification and invokes the object through one of its multiple entry points. While not directly related to fault management, this approach may serve as a method for implementing adaptivity.

#### 4. RELIABILITY MODELING

##### Overview and examples

V.R. Basili and B.T. Perricone, Software Errors and Complexity: An Empirical Investigation, *Comm. ACM* 27,1 (January 1984), 42-52.

An analysis of the distributions and relationships derived from change data collected during the development of several systems between August 1977 and May 1980. The information collected includes changes to the system and repair of detected errors. The major findings were that, first, the majority of detected errors were due to misunderstanding of the specifications (!) and, second, that the size of a module was not correlated to its proneness to errors.

V.R. Basili, Recent Advances in Software Measurement, *Proc. 12th Int. Conf. on Software Eng.* (March 1990), IEEE Comp. Soc. Press, 44-49.

An abstract of a talk given at the conference. The paper doesn't have much to say, but there is an extensive bibliography on software reliability measurement included.

M.W. Bush, Getting Started on Metrics -- Jet Propulsion Laboratory Productivity and Quality, *Proc. 12th Int. Conf. on Software Eng.* (March 1990), IEEE Comp. Soc. Press, 133-142.

A report on JPL's experience with collecting and analyzing data on the software development process. The collection process was part of a quality improvement effort and was intended to provide baselines for quality and productivity measures. Included in the paper are some figures on defects per lines of code in systems developed by JPL, IBM, and RADC.

K.C. Ferrara, S.J. Keene, and C. Lane, Software Reliability from a System Perspective, *Proc. Ann. Reliability and Maintainability Symp.* (January 1989), IEEE, 332-336.

A attempt to combine hardware and software reliability metrics to yield a system estimate. The paper documents a number of approaches to measuring reliability used in IBM to both evaluate and improve system products. An interesting observation is that in many cases the act of measuring development alone was seen to improve the process significantly.

A.L. Goel, Software Reliability Models, *IEEE Trans. on Software Eng.* SE-11,12 (December 1985), 1411-1423.

A survey (partially funded by RADC) of approaches to modeling reliability in software systems. The paper includes an assessment of the limitations and the assumptions of several models and proposes an approach for fitting a model to a problem. The examples use failure data from a C2 system developed at RADC. A classic paper.

D.I. Heimann, N. Mittal, and K.S. Trivedi, Dependability Modeling for Computer Systems, *1991 Proc. Annual Reliability and Maintainability Symp.* (January 1991), 120-128.

An up-to-date survey of measures and models for hardware dependability. The paper describes the three basic measures of dependability (availability, reliability, and task completion) and discusses how dependability analysis makes use of them. A discussion of parameter estimation for the models is also included.

S.J. Keene, Cost-Effective Software Quality, *Proc. Ann. Reliability and Maintainability Symp.* (January 1991), IEEE, 433-437.

Outlines the use of metrics in software development at IBM. Three methods of software reliability estimation -- precedent code, fault content modeling, and time domain modeling -- are described. Other techniques for improving quality both in software and in the development process are discussed.

H.D. Mills and P.B. Dyson, Using Metrics to Quantify Development, *IEEE Software* (March 1990), 15-16.

The guest editors' introduction to a special issue on software metrics. The whole problem of predicting software reliability is still controversial, partly because it is still, as the authors describe it, an "adolescent activity." Reliability measurement in hardware is well defined, but is much less so in software. The suggestion from the authors is to define a framework for measurement on a project-specific basis that allows you to get results in a form that can be used to monitor the development process and thus detect or predict problems.

J.D. Musa, The Measurement and Management of Software Reliability, *Proc. of the IEEE* 68,9 (September 1980), 1131-1143.

A survey and history of software reliability modeling from 1967-1980. Although older and not as complete as Musa's 1987 text (referenced below), this paper compares software and hardware reliability models and explains how the software models can be used to predict the behavior of an operational system based on information from its development and the code. A good summary of the early development of models.

J.D. Musa, A. Iannino, and K. Okumoto, *Software Reliability: Measurement, Prediction, Application*, McGraw-Hill (1987).

The single most useful reference for reliability modeling of software, both for its application and the underlying theory. The emphasis of this text is on the use of reliability measures for planning and evaluation development, but it may be that some of the measures presented here and elsewhere can be used for detection and adaptation in operational systems as well.

J.D. Musa, Quality Time: Faults, Failures, and a Metric Revolution, *IEEE Software* (March 1989), 85+.

A short editorial on the shift of quality measurements from the point of view of the developer (to whom faults are of utmost concern) to that of the user (who is worried about failures). "Quality" is a word that gets bandied about by lots of people, most of whom really don't know what they mean by it. From the user's point of view, quality means freedom from failure, and Musa argues that systems should be evaluated in terms of failures per hours of operation, rather than in faults per lines of code. He also suggests that an operational profile of a system (a list of the system functions and the proportion of time spent executing in each) is a useful tool for debugging a system as well as evaluating it. In particular, it can be used to reduce the failure rate in systems where usage of functions is nonuniform (i.e., make sure that the functions you use most work best).

### Modeling techniques

J. Arlat, K. Kanoun, and J.-C. Laprie, Dependability Modeling and Evaluation of Software Fault-Tolerant Systems, *IEEE Trans. on Computers* 39,4 (April 1990), 504-513.

Models that describe reliability and safety in software-fault tolerant architectures, including recovery blocks and N-version programming. The studies summarized here consider both independent and related faults and methods for combining the approaches to handle multiple faults. The results of the modeling indicate that related faults between the variants in multiversion architectures have a significant impact on safety, while the nesting of recovery-based architectures shows less improvement in reliability than other approaches. Another approach analyzed was the discarding of a variant that frequently disagrees with the others in a multiversion architecture; the study indicated that this reduction in functionality is always beneficial with respect to safety, and to reliability in the case where independent faults dominate.

W.K. Ehrlich, S.K. Lee, and R.H. Molisani, Applying Reliability Measurement: A Case Study, *IEEE Software* (March 1990), 56-64.

An example of using failure data to validate a reliability model based on historical data. The testbed was the AT&T telecommunication testing system RMS-D1.

J.K. Muppala, S.P. Woollet, and K.S. Trivedi, Real-Time Systems Performance in the Presence of Failures, *Computer* 24,5 (May 1991), 37-47.

Presents a unified model for evaluating performance, reliability/availability, and deadline violation based on stochastic reward nets and Markov chains. The model gives a response-time distribution for a queuing network. Throughputs and response time are modeled as reward rates. Examples include predicting the size of the Markov chains (useful since the response times decrease with the number of processors involved in a computation).

Y. Nakagawa and S. Hanata, An Error Complexity Model for Software Reliability Measurement, *Proc. 11th Int. Conf. on Software Eng.* (May 1989), IEEE Comp. Soc. Press, 230-236.

A time-dependent model for systems with complex errors developed by Nippon Telephone and Telegraph. The model predicts the number of remaining errors in software based on the number of detected errors and the ratio of complex to simple errors. The model has a better fit than other models on the data set presented, although more validation is needed before it can be used.

J.S. Ostroff and W.M. Wonham, Modeling, Specifying, and Verifying Real-Time Embedded Computing Systems, *Proc. Real-Time Sys. Symp.* (December 1987), IEEE Comp. Soc. Press, 124-132.

Proposes a framework for modeling and verifying hard real-time systems using extended-state machines. A specification language is described that was developed for use in this framework and provides an abstract operational and axiomatic semantics for programming constructs.

C.V. Ramamoorthy and F.B. Bastani, Software Reliability -- Status and Perspectives, *IEEE Trans. on Software Eng.* SE-8,4 (July 1982), 354-371.

A review of models based on the residual error size predicted and the testing process used. The paper also describes methods for estimating program reliability and the adequacy of test cases based on equivalence classes in the input domain. The method is expensive to use, but is useful for ultrareliable systems in which failure data is difficult to collect.

R.D. Schlichting, A Technique for Estimating Performance of Fault-Tolerant Programs, *IEEE Trans. on Software Eng.* SE-11, 6 (June 1985), 555-563.

Presents models for estimating the performance of fault-tolerant programs executing on fail-stop processors. The models use discrete-time Markov chains and z-transforms to derive a probability distribution for time to completion based on numbers of failures. The results are useful for predicting a program's ability to meet real-time deadlines in the presence of failures in the system.

Y.-B. Shieh, D. Ghosal, and S.K. Tripathi, Modeling of Fault-Tolerant Techniques in Distributed Systems, *Proc. 19th Int. Symp. on Fault-Tolerant Computing* (June 1989), IEEE Comp. Soc. Press, 167-174.

An evaluation of arbitrary versus planned checkpointing in both centralized and distributed systems. The checkpointing protocol is modeled using Petri nets. The results for the distributed scheme show that as the synchronization interval decreases, arbitrary checkpointing is faster than planned (coordinated) checkpointing due to the overhead of the processes taking their checkpoints within a given interval. In addition to the Petri net modeling, this paper is interesting as an example of adaptive specification, since it suggests that planned checkpointing is most effective as long as the synchronization interval remains above a specified threshold.

R.M. Smith, K.S. Trivedi, and A.V. Ramesh, Performability Analysis: Measures, an Algorithm, and a Case Study, *IEEE Trans. on Computers* 37,4 (April 1988), 406-417.

Models and algorithms for determining performability (a combined measure of performance and reliability) in heterogeneous distributed systems, both with and without repair. The systems are modeled as Markov chains with reward rates associated with each state. The results indicate that distributions of cumulative performance measures over finite intervals reveal behaviors not indicated by either steady-state models or the expected values alone, making such a measure an attractive possibility for an objective function.

M. Takahashi and Y. Kamayachi, An Empirical Study of a Model for Program Error Prediction, *Proc. 8th Int. Conf. on Software Eng.* (August 1985), IEEE Comp. Soc. Press, 330-336.

Presents a model to estimate the number of errors in a program prior to testing, using environmental and development factors identified through regression analysis. The results show that the frequency of specification change, the skill of the developer, and the volume of the design documentation contribute significantly to the number of errors in software and can be used to give a more accurate prediction of the reliability of software than size measures alone.

### Automated tools

R. Geist and K. Trivedi, Reliability Estimation of Fault-Tolerant Systems: Tools and Techniques, *Computer* 23,7 (July 1990), 52-61.

A review of tools and techniques for reliability estimation. Among the tools profiled is CARE, which is described in more detail in Chapter 9 of the Pradhan text (Volume II) referenced in the overview section.

A. Goyal et al., The System Availability Estimator, *Proc. 16th Int. Symp. on Fault-Tolerant Computing* (July 1986), IEEE Comp. Soc. Press, 84-89.

Description of the SAVE package, which is used to construct probabilistic models of availability and reliability. The system, intended to be used during system design and configuration, allows parameters to be defined separately for each component in the system. The parameters may include number of spares, the spares failure rate, the module that both operation and repair depend upon, the component failure rate (actual or probability), and a list of affected components.

J.D. Musa, Tools for Measuring Software Reliability, *IEEE Spectrum* 26,2 (February 1989), 39-42.

Some comments on programs that exist for computing software reliability using Musa's basic and logarithmic Poisson models, as described in his text. In general, the models are useful for project management (predicting release dates, etc.) but are not sufficiently precise for ultrareliable systems. An interesting sidebar is the description of an approach to "reliability engineering" that uses reliability models to help identify the problem to be solved and the information that should be collected throughout the system's life to verify its reliability.

## **5. ADDITIONAL SOURCES OF INFORMATION**

This bibliography is not intended to be exhaustive. Rather, it contains references that are particularly relevant to the issues discussed in the preceding report. More information on dependable distributed computing can be found from the sources listed below.

### **Annual conferences with a primary focus on dependability**

International Symposium on Fault-Tolerant Computing  
IEEE Symposium on Reliable Distributed Systems  
(formerly Symposium on Reliability in Distributed Software and Database Systems)  
Reliability and Maintainability Symposium  
Pacific Rim International Symposium on Fault-Tolerant Systems (Japan)  
International Symposium on Fault-Tolerant System Design (Bulgaria)

### **Other conferences with occasional papers on dependability**

ACM Symposium on Principles of Distributed Computing  
Real-Time Systems Symposium  
International Conference on Distributed Computing Systems  
International Conference on Software Engineering

### **Journals where articles on dependability often appear**

*IEEE Transactions on Reliability*  
*IEEE Transactions on Computers*  
*IEEE Transactions on Software Engineering*  
*IEEE Transactions on Parallel and Distributed Systems*  
*Communications of the ACM*  
*Computer*

(special issues on fault-tolerant computing were published in July 1990 and August 1984; other papers appear occasionally)

It is interesting to note that the August 1991 issue of *Byte* contains several articles relating to dependability, particularly mirrored disk systems for reliably maintaining data. This



movement into the microcomputer arena, as well as into the "popular" literature, could indicate a wider audience and increased interest in dependability issues.

**1991 USAF-RDL SUMMER RESEARCH PROGRAM  
FOR FACULTY AND GRADUATE STUDENTS**

**Sponsored by the  
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH**

***FINAL REPORT***

**ANALYSIS OF THE ELECTROMIGRATION INDUCED FAILURE  
IN THE VLSI INTERCONNECTION COMPONENTS AND  
THE MULTISECTION INTERCONNECTIONS**

**Prepared by :** Ashok K. Goel, Ph.D.  
Assistant Professor  
and  
Matthew M. Leipnitz  
Graduate Student

**Department :** Department of Electrical Engineering

**University :** Michigan Technological University

**Research Location :** Reliability Physics Laboratory  
RL/RBRP  
Rome Air Development Center  
Griffiss Air Force Base  
Rome, N.Y. 13441

**USAF Researcher :** Martin J. Walter

**Date :** July 17, 1991

**Contract Number :** F-49620-90-C-0076

# **ANALYSIS OF THE ELECTROMIGRATION-INDUCED FAILURE IN THE VLSI INTERCONNECTION COMPONENTS AND THE MULTISECTION INTERCONNECTIONS**

*Ashok K. Goel, Assistant Professor, and*

*Matthew M. Leipnitz, Graduate Student*

Department of Electrical Engineering  
Michigan Technological University  
Houghton, MI 49931

## **ABSTRACT**

We have carried out a first-order analysis of the electromigration induced failure effects in the various VLSI interconnection components including the multisection interconnections using the series model for failure mechanism. The Components include a straight interconnection segment, an interconnection bend, an interconnection step, an interconnection plug, an interconnection via, an interconnection overflow, a horizontal multisection interconnection, a vertical multisection interconnection, a mixed multisection interconnection, and a power/ground bus. First, by considering the effect of average flux density on the grain-boundary migration, we have reduced each interconnection component in to a series or series-parallel combination of straight segments. Then, for each of the components, we have investigated the dependence of the median-time-to-failure and the Log-normal standard deviation of the corresponding failure distribution on the various component parameters. The results can be utilized to choose optimum values of the component parameters for minimum probability of interconnection failure due to electromigration.

## 1. INTRODUCTION

Continuous advances in the field of VLSI are resulting in smaller integrated circuit chips having millions of interconnections that integrate the components on the chip. Interconnection failure caused by electromigration is one of the major factors responsible for lowering the effective lifetime of the chip [1-3]. Therefore, it is very important to understand the dependence of the electromigration induced interconnection failure on the various interconnection parameters. In fact, in the past, several studies have been dedicated to this effort [4-8].

In general, an interconnection line on an IC Chip consists of several components such as straight segments, bends, steps, plugs and vias. In addition, there are power and ground buses serving several logic gates on the chip. For submicron width interconnection lines, there can be sections along the line length suffering from material overflows. In this report, for the first time, we also introduce a multisection interconnection which can be designed in three possible configurations: horizontal, vertical and mixed. A multisection intersection differs from a generally employed interconnection in that a driver and its load are connected by more than one interconnection line thus providing more than one path for the current/voltage signal to flow.

In this report, we present an analysis of the electromigration-induced failure in each of the interconnection components listed above. First, using the effects of the average flux density on the grain boundary migration in the interconnections, we have derived expressions for the effective lengths, widths and thicknesses of the straight segments equivalent to each of the components. Then we have used the series model of failure mechanisms in the interconnections [9] to determine the series or series-parallel combinations of straight segments equivalent to each interconnection component. Finally, we have studied the dependence of the electromigration-induced Median-Time-to-Failure (MTF) and the standard deviation of the corresponding lognormal failure distribution ( $\sigma$ ) on the various parameters of each interconnection component.

## 2. REDUCTION OF COMPONENTS INTO STRAIGHT SEGMENTS

First, we have analyzed a straight interconnection segment, shown in Figure 1(a), of length  $L$ , width  $W$  and thickness  $T$  carrying a current  $I$  at a given temperature. Then, by considering the effects of the average flux density on the grain boundary migration in each interconnection component, we have reduced it to a series-parallel combination of equivalent straight interconnection segments. The average flux density in a component was determined by using the interconnection current  $I$  and the average cross sectional area throughout the component.

The additional area in an interconnection bend of angle  $\theta_B$ , shown shaded in Figure 1(b), was found equivalent to a straight segment of length  $L_B$  and width  $W_B$  given by the expressions:

$$L_B = \frac{\pi W (180 - \theta_B)}{360}$$

$$W_B = \frac{W^2 (1 + \sqrt{\tan(\theta_B/2)})}{W + L_B}$$

For a bend angle of 90 degrees, these expressions yield values in agreement with those derived by Frost and Poole [9].

An interconnection line of length  $L$ , width  $W$  and thickness  $T$  having a single step of height  $H$  and angle  $\theta_S$ , shown in Figure 1(c), is equivalent to three straight segments each of width  $W$ , lengths  $L_{S_1}$ ,  $L_{S_2}$  and  $L_{S_3}$ , and thicknesses  $T_{S_1}$ ,  $T_{S_2}$  and  $T_{S_3}$ , respectively given by the expressions:

$$L_{S_1} = L + \frac{H}{\tan \theta_S}$$

$$T_{S_1} = T$$

$$L_{S_2} = T \cos \theta_S + \frac{H}{\sin \theta_S}$$

$$T_{S_2} = T \cos \theta_S$$

$$L_{S_3} = \frac{\pi T (180 - \theta_S) (1 - \cos \theta_S)}{720}$$

$$T_{S_3} = \frac{T^2 (1 + \cos^2 \theta_S - \sin \theta_S \cos \theta_S)}{2 L_{S_3} + T (1 - \cos \theta_S)}$$

Two straight sections of an interconnection line of total length  $L$ , width  $W$  and thickness  $T$  joined by a single plug of length  $H$  and square dimension  $W_P$ , shown in Figure 1(d), is equivalent to three straight segments of lengths  $L_{P_1}$ ,  $L_{P_2}$  and  $L_{P_3}$ , widths  $W_{P_1}$ ,  $W_{P_2}$  and  $W_{P_3}$ , and thicknesses  $T_{P_1}$ ,  $T_{P_2}$  and  $T_{P_3}$  respectively given by the expressions:

$$L_{P_1} = L - W_P$$

$$W_{P_1} = W$$

$$T_{P_1} = T$$

$$L_{P_2} = H$$

$$W_{P_2} = W_P$$

$$T_{P_2} = W_P$$

$$L_{P_3} = \frac{\pi}{8} (T + W_P)$$

$$W_{P_3} = W$$

$$T_{P_3} = \frac{T (W + W_P)}{L_{P_3}}$$

An interconnection line of length  $L$ , width  $W$  and thickness  $T$  having a length  $L_O$  suffering from overflow (top and end views are shown schematically in Figure 1(e)), is equivalent to two straight segments of lengths  $L_{O_1}$  and  $L_{O_2}$ , widths  $W_{O_1}$  and  $W_{O_2}$ , and thicknesses  $T_{O_1}$ ,  $T_{O_2}$  given by the expressions:

$$L_{O_1} = L - L_O$$

$$T_{O_1} = T$$

$$W_{O_1} = W$$

$$L_{O_2} = L_O$$

$$T_{O_2} = \frac{-W + \sqrt{W^2 + 4 W T}}{2}$$

$$W_{O_2} = \frac{W T}{T_{O_2}}$$

Two straight sections of an interconnection line of total length  $L$ , width  $W$  and thickness  $T$  joined by a via of height  $H$ , width  $W_V$  and angle  $\theta_V$ , shown in Figure 1(f), is equivalent to four straight segments each of width  $W$ , lengths  $L_{V_1}$ ,  $L_{V_2}$ ,  $L_{V_3}$  and  $L_{V_4}$ , and thicknesses  $T_{V_1}$ ,  $T_{V_2}$ ,  $T_{V_3}$  and  $T_{V_4}$  respectively given by the expressions:

$$L_{V_1} = \frac{\theta_V \pi T (1 + \cos(\theta_V))}{720}$$

$$T_{V_1} = \frac{T^2 [1 + \cos\theta_V \sin\theta_V + \cos^2 \theta_V]}{2 L_{V_1} + T [1 + \cos\theta_V]}$$

$$L_{V_2} = \frac{H}{\sin\theta_V} - T \sin\theta_V$$

$$T_{V_2} = T \cos\theta_V$$

$$L_{V_3} = \frac{T}{2 \sin \left[ \tan^{-1} \left[ \frac{T}{W_V} \right] \right]}$$

$$T_{V_3} = \frac{\frac{T^2}{2 \tan\theta_V} + T \left[ W_V - \frac{T}{\tan\theta_V} \right]}{L_{V_3}}$$

$$L_{V_4} = L - W_V - \frac{H}{\tan\theta_V}$$

$$T_{V_4} = T$$

A horizontal multisection interconnection of length  $L$  consists of a parallel combination of  $N$  interconnections each of length  $L$ , width  $W$  and thickness  $T$  placed between two rectangular pads each of length  $L_{pad}$  and width  $W_{pad} = (2N - 1)W$  on each side which are in turn connected to the interconnection driver and its load. A schematic diagram of the top view of a horizontal multisection interconnection is shown in Figure 1(g).

A vertical multisection interconnection of Length  $L$  consists of a parallel combination of  $N$  interconnections one of which is printed on top of the substrate while the others are embedded in the substrate exactly below the top section. Each section is of length  $L$ , width  $W$  and thickness  $T$ . The sections are connected to each other at the ends by conducting plugs each of length  $L_P$ . A schematic diagram of the side view of a vertical multisection interconnection is shown in Figure 1(h). A mixed multisection interconnection is formed by mixing the horizontal and vertical multisection interconnections i.e. it has a few sections printed on top of the substrate in addition to a few sections embedded in the substrate.

As shown in Figure 1(i), a power or ground bus serving  $N_g$  gates on the integrated circuit chip was modelled as a series combination of  $N$  straight segments carrying currents equal to  $I, 2I, 3I, \dots, N_g I$  where  $I$  is the current in each gate.

### 3. CALCULATION OF MTF AND LOG-NORMAL STANDARD DEVIATION

First, for a basic conductor element of length  $10\mu m$ , the median-time-to-failure was found by using the expression [9]:

$$MTF = 1,523.0 \left[ \frac{W T}{I \times 10^5} \right]^n \left[ W - 3.07 + \frac{11.63}{W^{1.7}} \right] e^{10,740.74 E_a / T_K}$$



where  $I$  is the interconnection current in mA,  $n$  is the current density exponent,  $E_a$  is the activation energy of the interconnection material in eV,  $T_K$  is the temperature in Kelvins,  $W$  is the interconnection width in  $\mu m$  and  $T$  is the interconnection thickness in  $\mu m$ . Then, as a first approximation, the median-time-to-failure of a series combination of  $N$  elements ( $MTF_s$ ) was found by using the expression:

$$\frac{1}{MTF_s} = \frac{1}{MTF_1} + \frac{1}{MTF_2} + \dots + \frac{1}{MTF_N}$$

whereas that of a parallel combination of  $N$  elements ( $MTF_p$ ) was found by using the expression:

$$MTF_p = MTF_1 + MTF_2 + \dots + MTF_N$$

The lognormal standard deviation ( $\sigma$ ) of a basic conductor element of width  $W$  ( $\mu m$ ) was given by [9]:

$$\sigma(W) = \frac{2.192}{W^{2.625}} + 0.787$$

Then, for a straight segment of length  $L$ , it was calculated by using the expression:

$$\sigma_n = \sigma n^{-0.304}$$

where

$$n = \frac{L(\mu m)}{10}$$

#### 4. THE PROGRAMS "EMVIC", "EMVIC-2" and "EMGRAPH"

To date, we have developed three programs called EMVIC, EMVIC-2 and EMGRAPH. Each of the programs is interactive and extremely user-friendly. EMVIC and EMVIC-2 are written in FORTRAN-77 while the graphics program called EMGRAPH is written in C.

The program EMVIC can be used to determine the MTF and Lognormal standard deviation of a straight interconnection segment (sis), interconnection bend (ib), interconnection step (is), interconnection plug (ip), interconnection via (iv), interconnection overflow (io), horizontal multisection interconnection (hmsi), vertical multisection interconnection (vmsi), mixed multisection interconnection (mmsi) and a power/ground bus (pgbus). (The symbols in parentheses were used to name the data files as explained below.) First, the user uses the default values or chooses his/her own values for the several parameters of any component listed above. For a straight segment, the parameters include its length (il), width (iw), thickness (it), temperature (temp), current (curr), current density exponent (cde) and its material's activation energy (iae). In addition to these parameters, the other components are defined by the additional parameters listed below:

**INTERCONNECTION BEND:** Bend Angle (ba)

**INTERCONNECTION STEP:** Step Height (sh), Step Angle (sa)

**INTERCONNECTION PLUG:** Plug Length (pl), Square Plug Dimension (pd), Plug Material Activation Energy, Lower Level Material Activation Energy

**INTERCONNECTION VIA:** Via Height (vh), Via Width (vw), Via Angle (va), Lower Level Material Activation Energy

**INTERCONNECTION OVERFLOW:**

Overflow Length (ol)

**HORIZONTAL MULTISECTION INTERCONNECTION:**

Number of Horizontal Sections (nhs), Source/Sink Pad Lengths (spl)

**VERTICAL MULTISECTION INTERCONNECTION:**

Number of Vertical Sections (nvs), Vertical Plug

Lengths (vpl), Plug Material Activation Energy

**MIXED MULTISECTION INTERCONNECTION:**

Number of Horizontal Sections (nhs), Number of Embedded Vertical Sections (nvs), Vertical Plug Lengths (vpl), Source/Sink Pad Lengths (spl), Plug Material Activation Energy

**POWER OR GROUND BUS:** Number of Gates Served by the Bus (ng), Current in Each Gate (gc)

After the user defines the component, EMVIC calculates the MTF and  $\sigma$  for it and displays these values on the screen. The user can choose to write the simulation results on an output file called EMVIC.OUT.

The program EMVIC-2 incorporates and extends the program EMVIC in the sense that it allows the user to study the dependence of MTF and  $\sigma$  for any component on its several parameters. First, the user can define the reference component. Then he/she can select a variable parameter and choose its lowest and highest values for the dependence analysis. The number of points at which the analysis is to be carried out can also be selected by the user. Then, EMVIC-2 calculates MTF and  $\sigma$  at evenly distributed points in the range of analysis and writes the results on an output file called EMVIC-2.OUT. It also creates data files named COMPONENT-PARAMETER.DAT one for each parameter of every component analyzed by the user. The values of COMPONENT and PARAMETER are the symbols enclosed in the parentheses above. These data files are written in a format required by the graphics program EMGRAPH. At this point, the user can use EMGRAPH to plot the results of any of the above dependence studies. EMGRAPH also allows the user to change the appearance of the plot in order to create custom plots. These plots can then be sent to a printer. The source codes of EMVIC-2 and EMGRAPH contain nearly 6,000 and 1,000 lines, respectively. The flow chart of EMVIC-2 is shown in Figure 2.

## 5. SIMULATION RESULTS

The program EMVIC-2 has been used to study the dependence of MTF and  $\sigma$  on the various parameters of each interconnection component. In the following results, current density exponent was set at 1.0. (Due to length restriction on this report, only a few results will be included here.)

First, for a straight interconnection segment, the dependence of MTF and  $\sigma$  on the segment width in the range 0.5-5  $\mu m$  is shown in Figure 3. The relatively sharp increase in MTF and  $\sigma$  for widths less than nearly 2  $\mu m$  is due to the so-called "bamboo" effect [4,5]. For an interconnection bend, the dependence of MTF and  $\sigma$  on the bend angle in the range 10-150 degrees is shown in Figure 4. This figure shows that a bend angle of nearly 50 degrees results in the lowest value of MTF. For an interconnection step, the dependence of MTF and  $\sigma$  on the step angle in the range 90-160 degrees is shown in Figure 5. This figure shows that MTF decreases rapidly as the step angle approaches 90 degrees. This is because of the gradual thinning of the material at the step. For a horizontal multisection interconnection, the dependence of MTF and  $\sigma$  on the number of horizontal sections in the range 1-5 is shown in Figure 6. This figure shows that MTF varies nearly as  $n^2$  where  $n$  is the number of sections. This is because the current density in each section is nearly  $(1/n)$  of that in the original single-section interconnection and further because all sections must fail before the interconnection fails completely. Similar dependence on the number of vertical sections was observed for the vertical multisection interconnection as shown in Figure 7. However, compared to the horizontal configuration, vertical multisection interconnection offers the advantage that it does not require any additional space on the chip. For a 1,000  $\mu m$  long power or ground bus serving 100 identical gates, the dependence of MTF and  $\sigma$  on the current in each gate in the range 0.1-1 mA is shown in Figure 8. This figure shows that increasing the gate currents results in lower values of MTF for the bus, as expected.

## 6. SUMMARY AND CONCLUSIONS

To summarize, we have carried out a first-order analysis of the electromigration induced failure effects in the several interconnection components. First, each component has been reduced into a series-parallel combination of equivalent straight interconnection segments and, then, the series model of failure mechanism has been used to determine the MTF and the corresponding lognormal standard deviation ( $\sigma$ ) for each component. The algorithms have been used to study the dependence of MTF and  $\sigma$  on the various parameters of each interconnection component. Though, due to the approximations inherent in the series model, the analysis presented in this report gives approximate results, yet these can be used to draw important conclusions regarding the optimization of the various interconnection components.

## ACKNOWLEDGEMENTS

First, we like to thank the United States Air Force Systems Command, the Air Force Office of Scientific Research and the Reliability Physics Laboratory of the Rome Air Development Center at the Griffis Air Force Base for sponsoring this research. We also like to thank several individuals who helped in making this experience truly rewarding and enriching for us. First, we are grateful to our technical focal point Mr. Martin Walter for his constant encouragement, several useful discussions and for providing a very constructive and enjoyable environment for carrying out this research. We are also grateful to Robert Hillman, Alfred Tamburrino and Mark Pronobis for taking interest in our work and for their helpful suggestions. We also like to thank Gary Moore and other staff members of the Research and Development Laboratories for their help with the several managerial aspects of this program.

## BIBLIOGRAPHY

- [1] J.R. Black, "Physics of Electromigration," Proc. 12th Annual Reliab. Phys. Symp., pp. 142-149, 1974.
- [2] P.B. Gbate, "Electromigration-Induced Failures in VLSI Interconnects," Proc. 20th Int. Reliab. Phys. Symp., pp. 292-299, 1982.
- [3] D.J. LaCombe and E.L. Parks, "The Distribution of Electromigration Failures," Proc. 24th Int. Reliab. Phys. Symp., pp. 1-6, 1986.
- [4] E. Kinsborn, "A Model for the Width Dependence of Electromigration Lifetimes in Aluminum Thin-Film Stripes," Appl. Phys. Lett., Vol. 36, pp. 968-970, 1980.
- [5] S. Vaidya, T.T. Sheng and A.K. Sinha, "Linewidth Dependence of Electromigration in Evaporated Al-0.5% Cu," Appl. Phys. Lett., Vol. 36, pp. 464-466, 1980.
- [6] J. Cho and C.V. Thompson, "Grain Size Dependence of Electromigration-Induced Failures in Narrow Interconnects," Appl. Phys. Lett., Vol. 54, No. 25, pp. 2577-2579, 1989.
- [7] Y.E. Strausser, B.L. Euzent, R.C. Smith, B.M. Tracy and K.Wu, "The Effect of Metal Film Topography and Lithography on Grain Size Distributions and on Electromigration Performance," Proc. Int. Reliab. Phys. Symp., pp. 140-144, 1987.
- [8] A.S. Oates, "Step Spacing Effects on Electromigration," Proc. Int. Reliab. Phys. Symp., pp. 20-24, 1990.
- [9] D.F. Frost and K.F. Poole, "A Method for Predicting VLSI-Device Reliability Using Series Models for Failure Mechanisms," IEEE Trans. Reliab., Vol. R-36, No. 2, pp. 234-242, 1987.

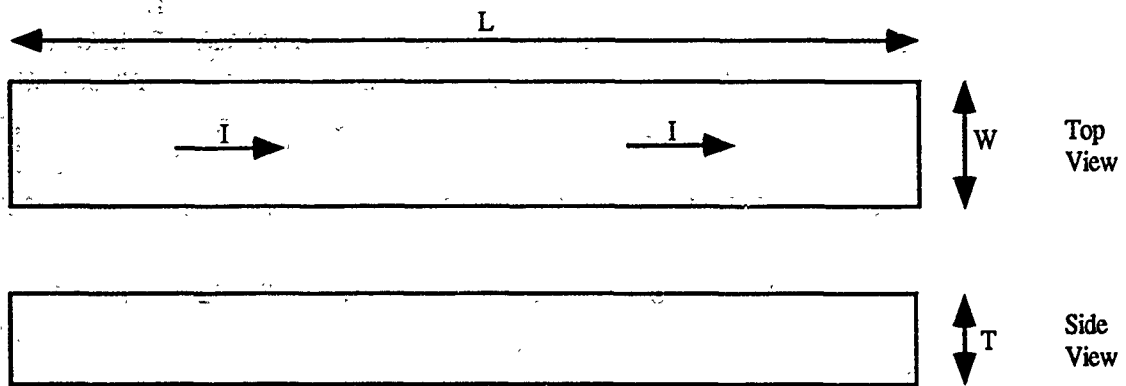


Figure 1(a): Schematic diagram of a Straight Interconnection Segment

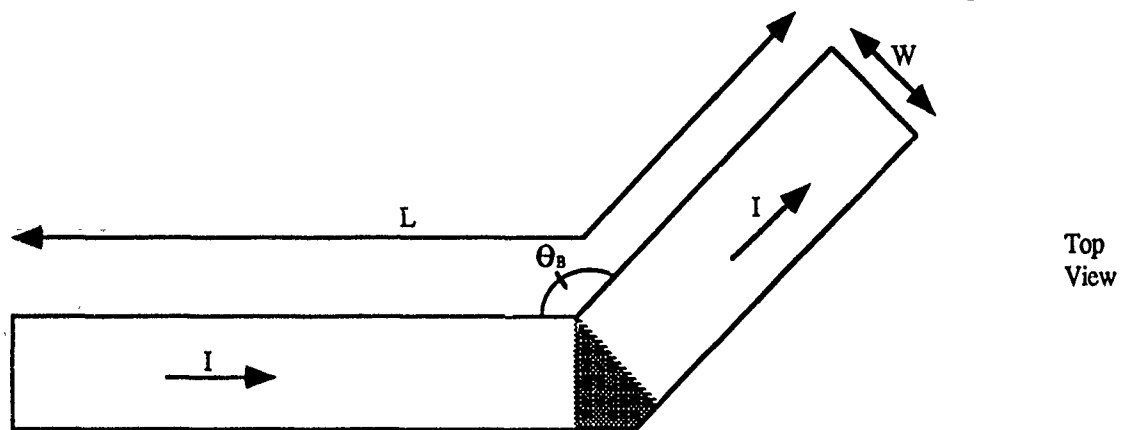


Figure 1(b): Schematic diagram of an Interconnection Bend

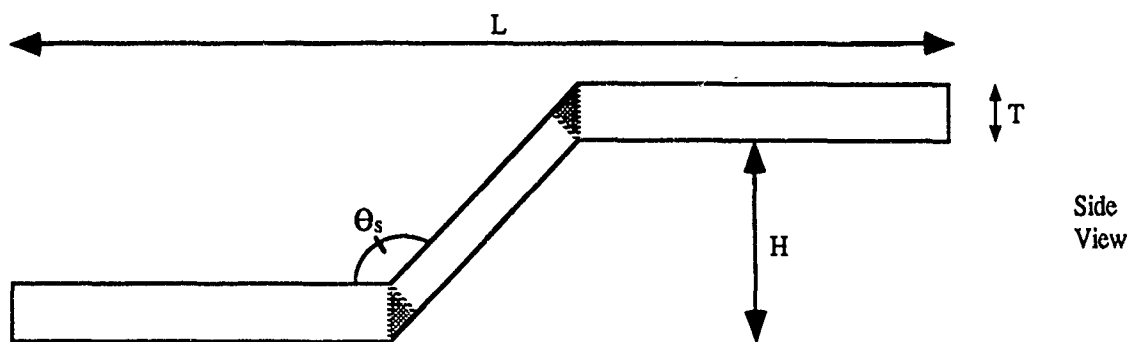


Figure 1(c) Schematic diagram of an Interconnection Step

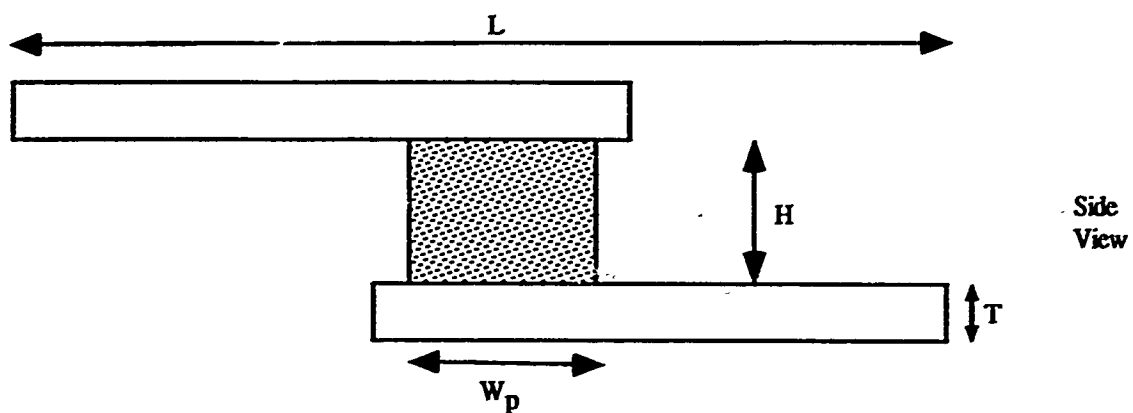


Figure 1(d): Schematic diagram of an Interconnection Plug

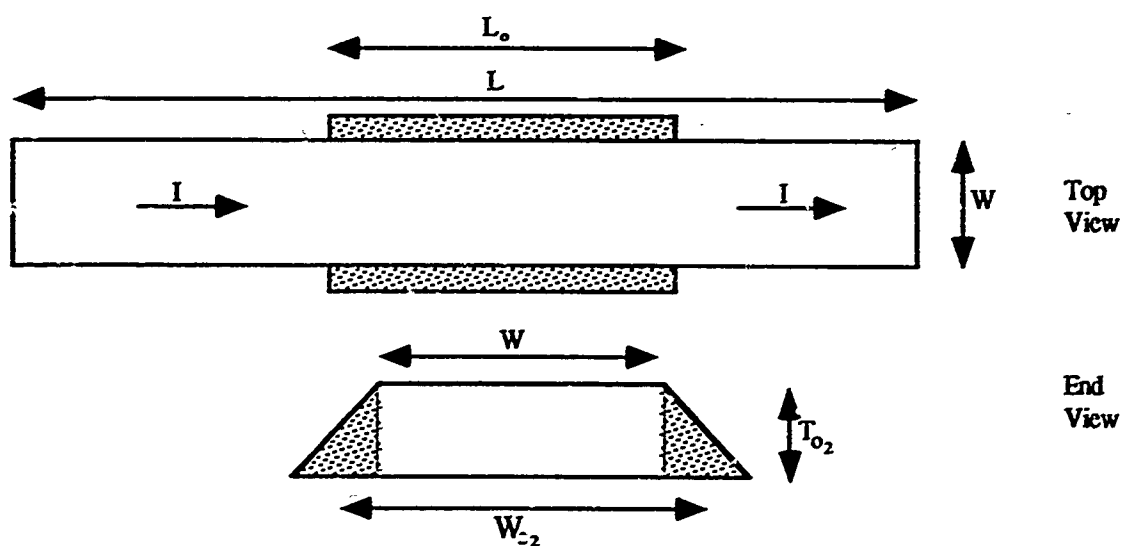


Figure 1(e): Schematic diagram of an Interconnection Overflow

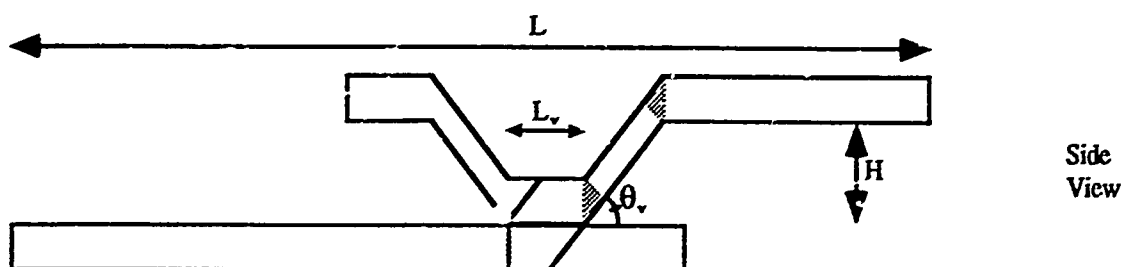


Figure 1(f): Schematic diagram of an Interconnection Via



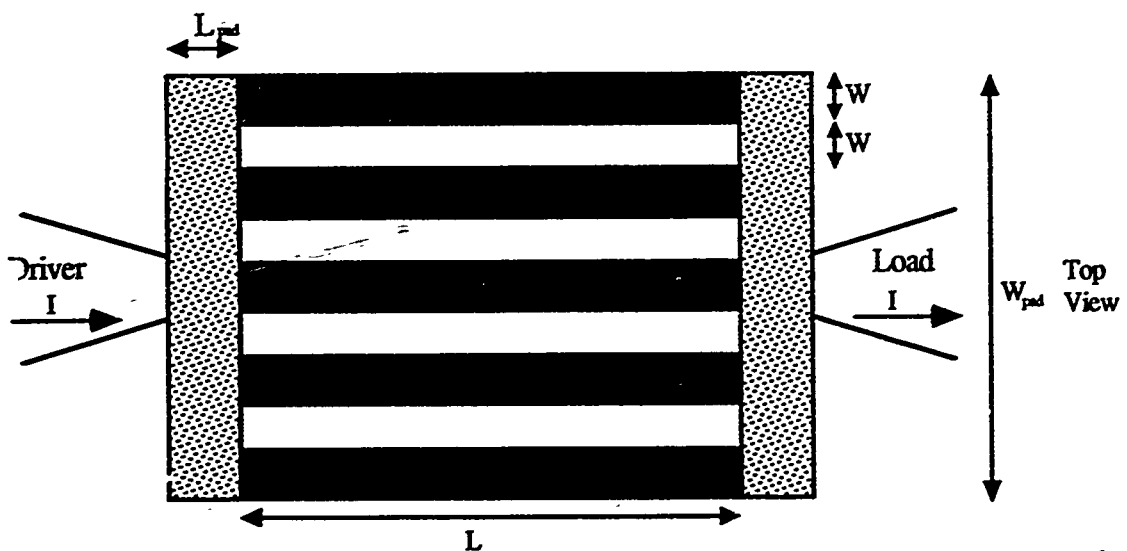


Figure 1(g): Schematic diagram of a Horizontal Multisection Interconnection

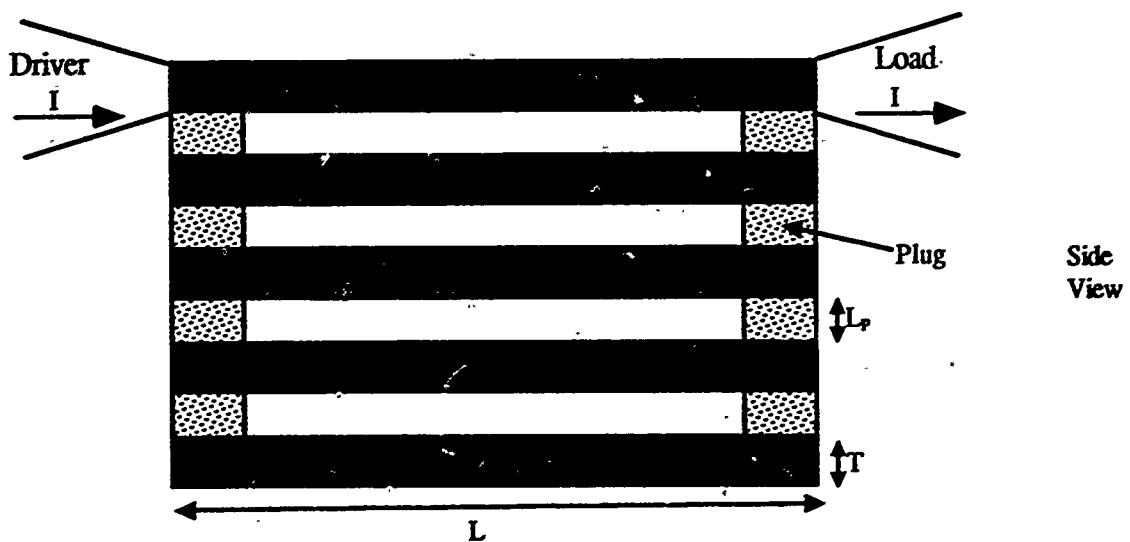


Figure 1(h): Schematic diagram of a Vertical Multisection Interconnection

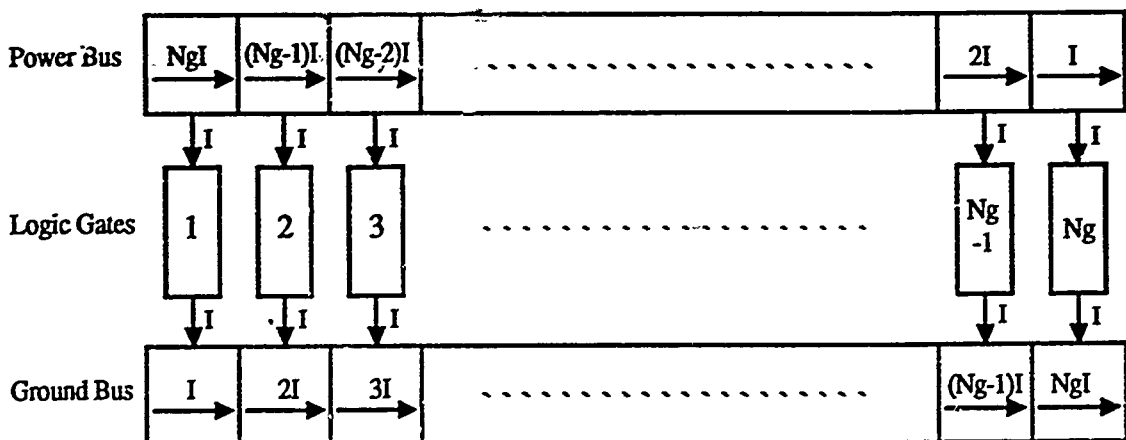
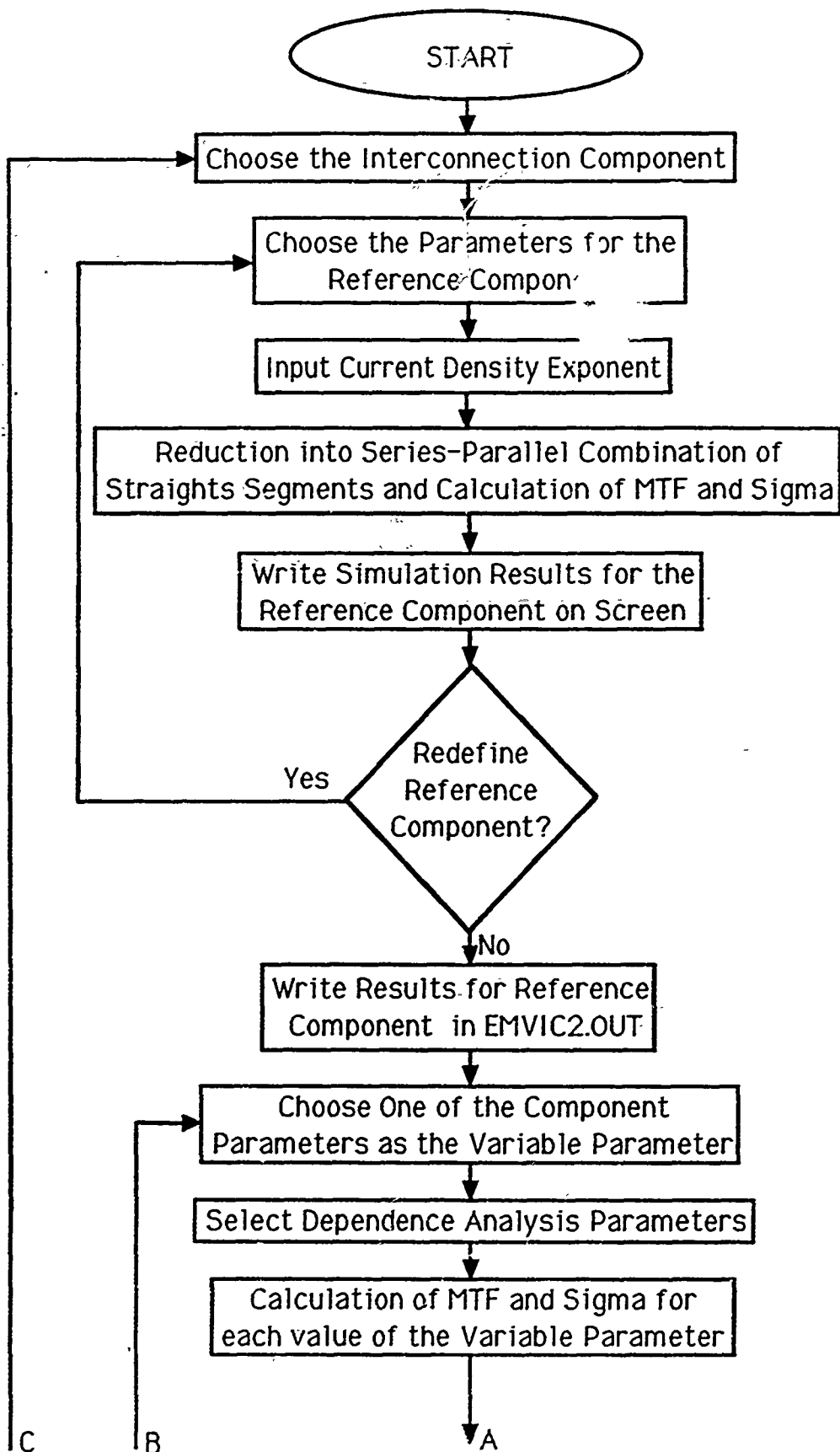
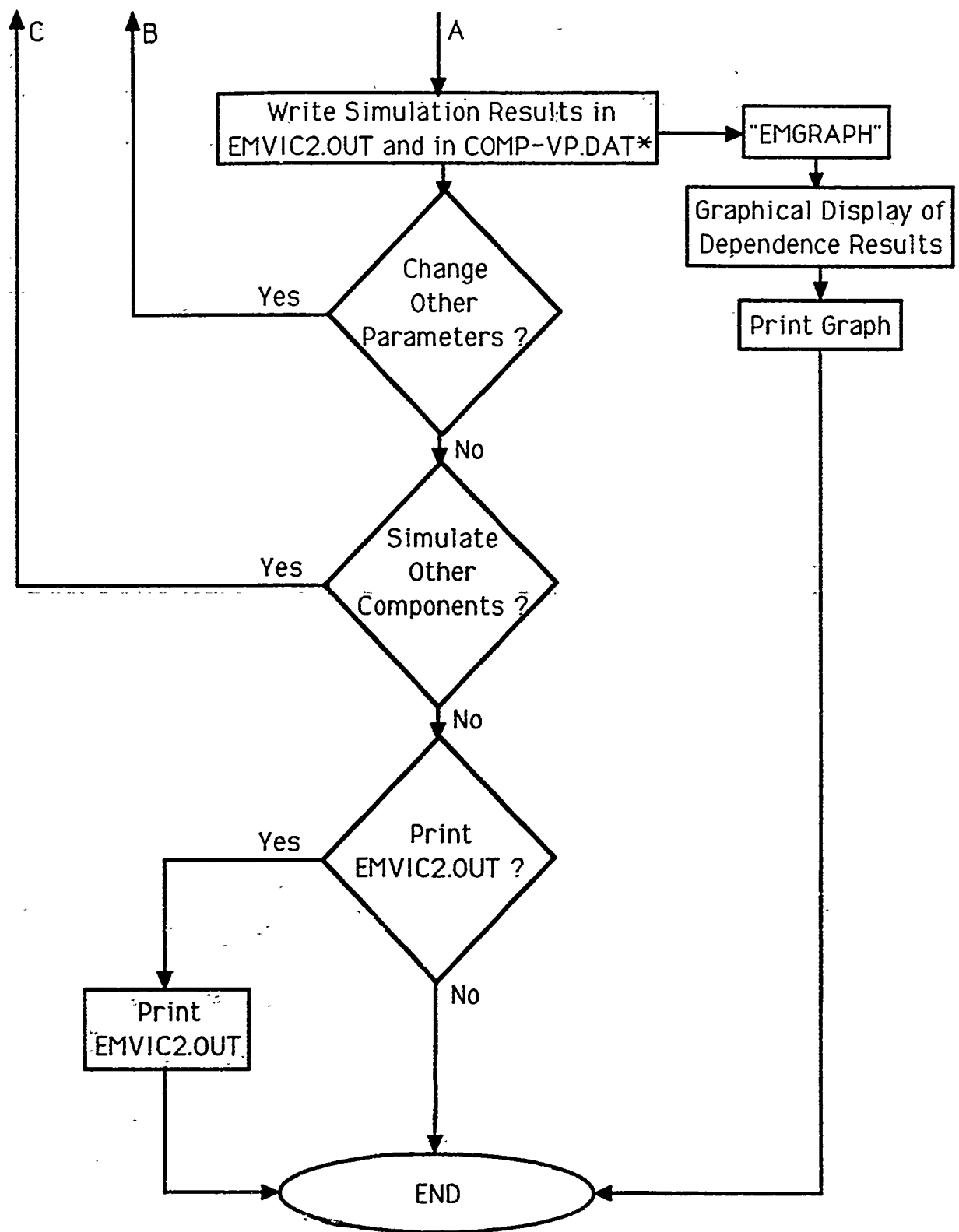


Figure 1(i): Schematic diagram of the Power and Ground Buses Serving  $N_g$  Gates on the Chip

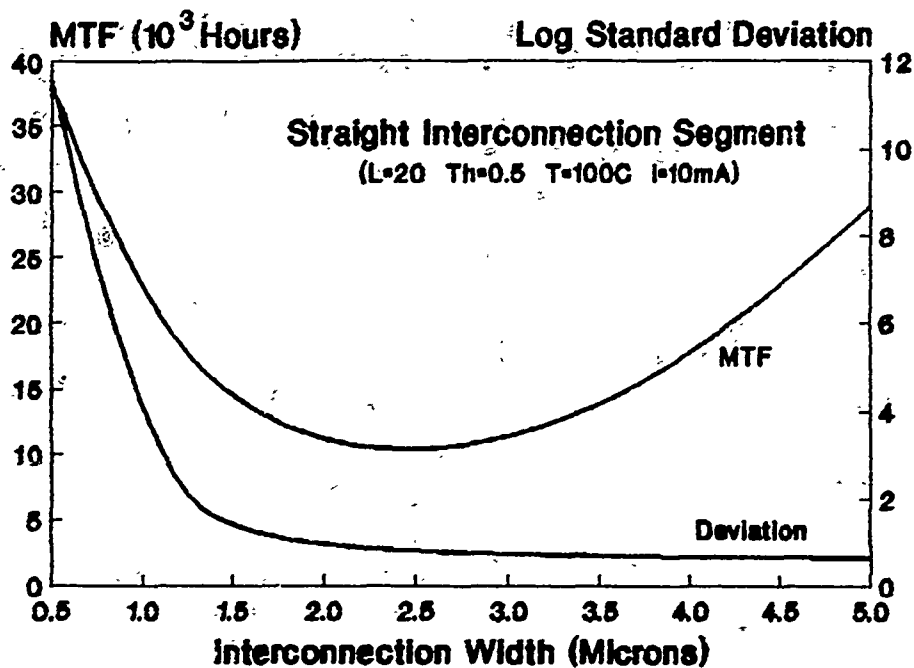


(Continued on Next Page)

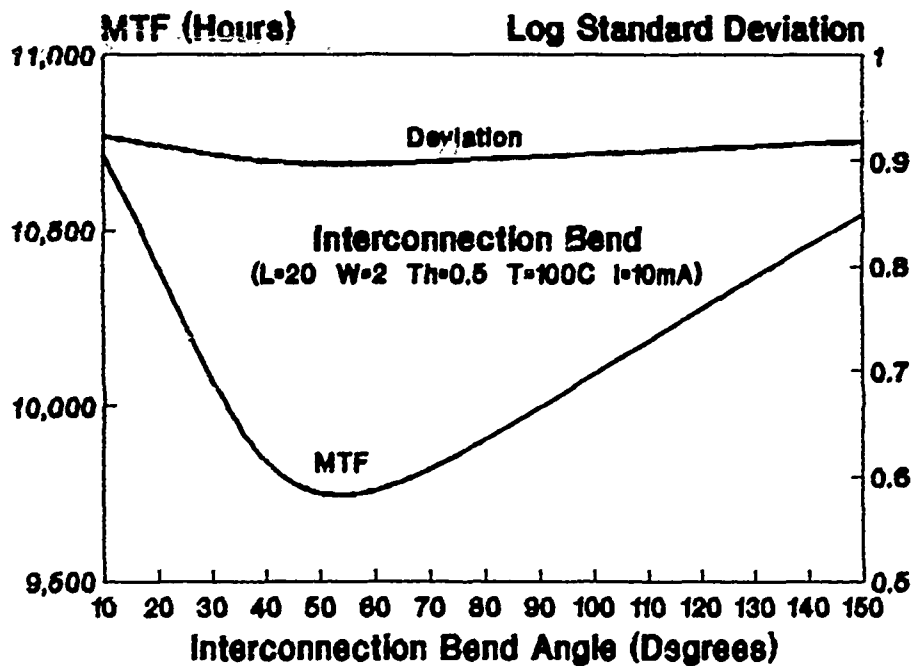


(\*NOTE: The file COMP-VP.DAT stands for Component Name-Variable Parameter Name.DAT)

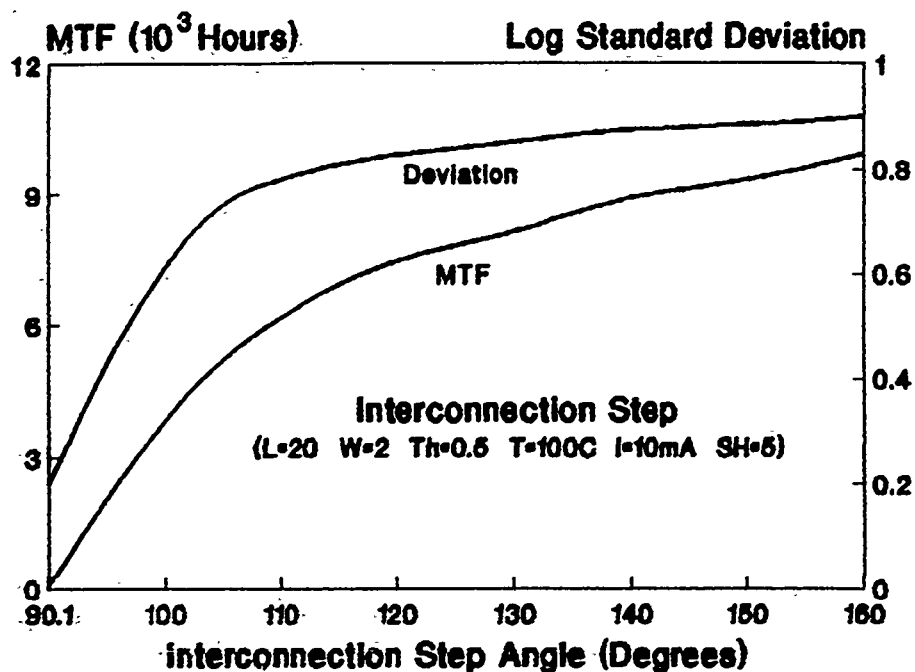
Fig. 2: Flow Chart of EMVIC-2



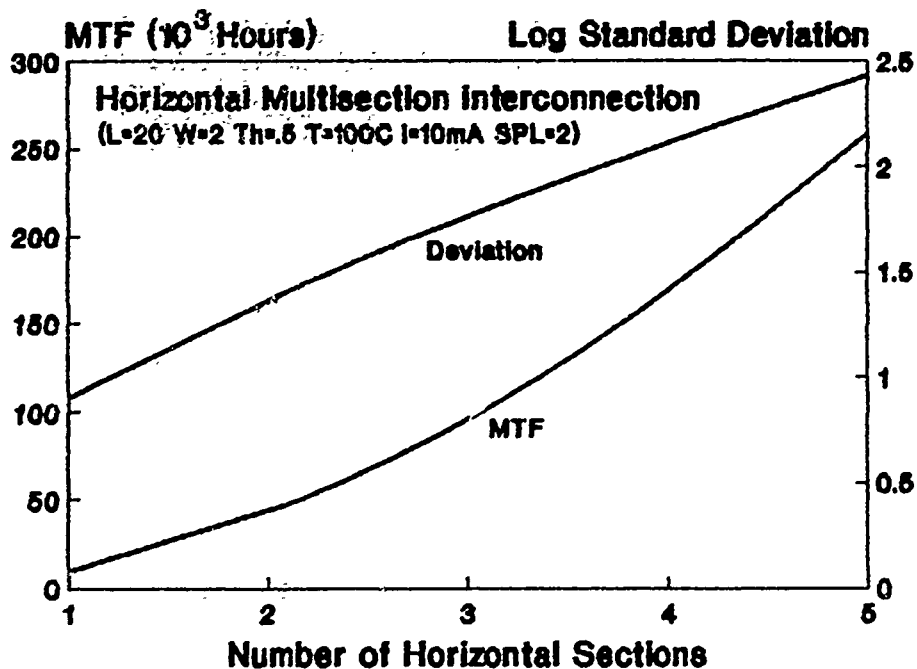
**Figure 3: MTF and Deviation vs. Width for SIS**



**Figure 4: MTF and Deviation vs. Angle for a Bend**



**Figure 5: MTF and Deviation vs. Angle for a Step**



**Figure 6: MTF and Sigma vs. No. of Sections for HMSI**

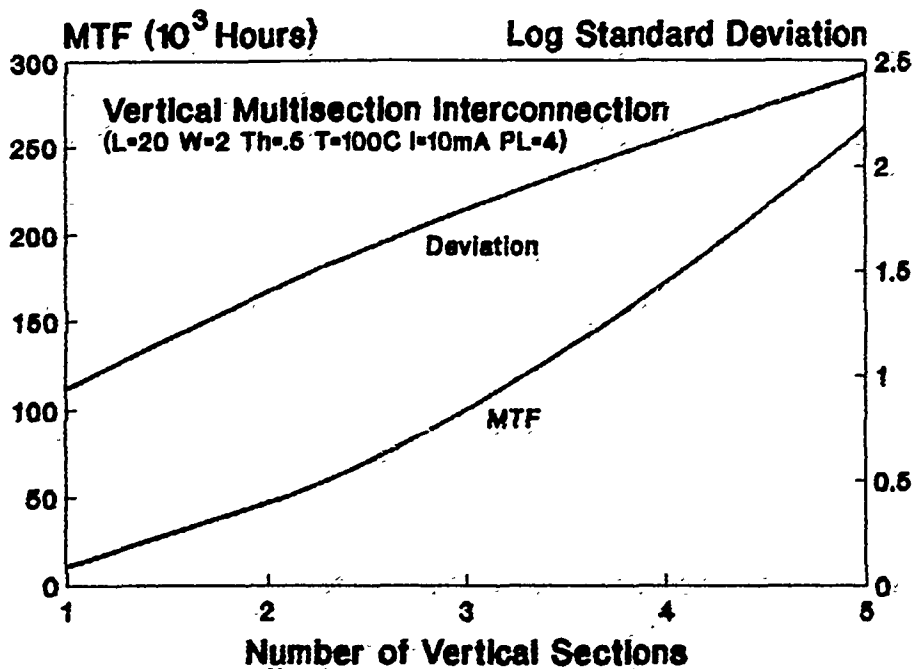


Figure 7: MTF and Sigma vs. No. of Sections for VMSI

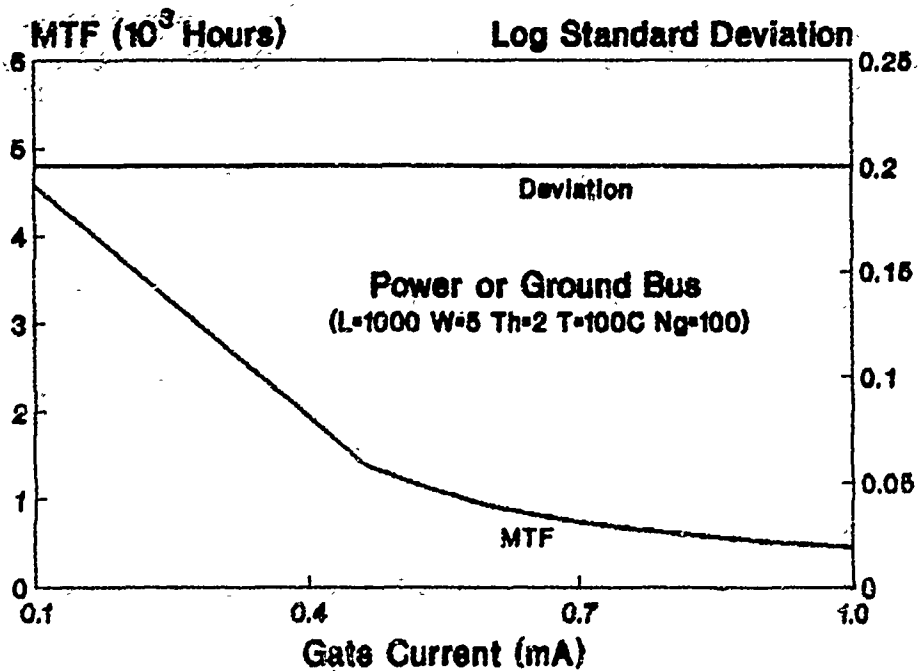


Figure 8: MTF and Sigma vs. Gate Current for a P/Q Bus

# FINAL REPORT SUMMER 1991

*by*

**Philipp Kornreich  
Employee No. 86**

**AFOSR SUMMER FACULTY RESEARCH PROGRAM**

**DEPARTMENT OF ELECTRICAL AND COMPUTER  
ENGINEERING**

**SYRACUSE UNIVERSITY  
SYRACUSE, N. Y., 13244  
Tel. No.: (315) 443 4447  
Fax. No.: (315) 443 2583**

# 1 INTRODUCTION

My original plan was to work this summer on optical wave guides and an optical Tunneling AND gate. The tunneling devices have to be fabricated completely in vacuum. A robot mask and substrate changer for a vacuum system at Syracuse University Micro-Electronics Laboratory (SUMEL) that is to be used for the fabrication of the tunneling devices was to be fabricated in the Rome Laboratory (RL) machine shop.

The robot mask changer was designed and it is now being fabricated in the RL machine shop. Preliminary test have been performed indicating that a light sensitive tunneling diode can be fabricated. Indeed, the continuation of this work using the robot mask changer is the subject of my follow on proposal.

While I was waiting for the robot mask changer I worked on two other projects. We tested a liquid crystal Spatial Light Modulator (SLM). and a binary lens for replicating light beams. This work is described below.

# 2 PHOTONIC TUNNELING AND GATE

The Photonic Tunneling AND Gate consist of two photosensitive tunneling diodes connected electrically connected in series. The individual diodes only conduct when illuminated. Since the conduction process in the diodes is due to the photo assisted tunneling effect the response of these devices should be exceedingly fast.

The device consists of a bottom aluminum film. This film is covered with a 50 Å thick aluminum oxide film. This oxide film form the tunnel barrier of the device. The edges of the bottom film are covered with a protective SiO<sub>2</sub> layer to prevent shorting of the top and bottom metal layers. The oxide layer is covered with a semitransparent aluminum film. A thicker aluminum is deposited over the semitransparent aluminum film between the bottom metal layer and the contact for the top metal. Both metal contacts are covered with a copper film. All these films have to be patterned. The patterning is achieved by depositing the various films through silicon wafer masks. These masks are fabricated by etching a crude pattern of the particular mask from the bottom of the wafer penetrating to within a few μm's of the top face of the wafer is. The top surface is etched in the desired pattern. This allows very accurate patterning.



The various silicon wafer masks and glass substrates can be manipulated by the mask and substrate changer robot.

Since the oxide layer is only 50 Å thick it is necessary for the bottom aluminum layer to be exceedingly smooth. Indeed, aluminum films deposited on glass in the SUMEL are smooth on a scale of less than about a few hundred Å. The aluminum films were analyzed with a Scanning Electron Microscope (SEM) at the RL. The SEM photograph showed that the aluminum films were smooth to the limit of the resolution of the SEM which is of the order of a few hundred Å.

### **3 LIQUID CRYSTAL LIGHT MODULATOR**

#### **3a INTRODUCTION**

A Spatial Light Modulator (SLM) is a device designed for computer controlled modulation of laser light on an individual element, or pixel, basis. SLM fabrication is conventionally done in two-dimensional arrays to provide sufficient area to cover typical laser beam dimensions (on the order of one mm<sup>2</sup>). The primary component of the SLM consists of an array that is comprised of two glass plates separated by a thin film of liquid crystal material. A reflective chromium mask creates 100 liquid crystal elements (pixels) in a 10x10 array pattern. Transparent electrical contacts are deposited on both the left and right side plates of the liquid crystal cell. One set of contacts is patterned and other is a common electrode. A polarizer is provided on the output side of the device. By applying a voltage across the contacts, an electric field is produced across the molecules in the liquid crystal which orient them in one of two preferred states that differ in alignment by 45 degrees. The state in which the molecules are aligned can be changed by switching the polarity of the electrical contacts. The array is illuminated with polarized coherent light. The output polarizer is arranged to block all light when there is no bias applied across the device. Applying a voltage across any cell of the SLM changes the alignment of the liquid crystal molecules which in turn rotates the polarization so that light is transmitted through the output polarizer.

The contact switching voltage is generated by a computer. If polarized light is incident on the array when the molecules in a liquid crystal element are aligned, the element functions as a half-wave plate. A half-wave plate rotates the polarization of the incident light through an angle twice that formed by the incoming polarization vector and the molecular state of orientation (optic axis). When an element is illuminated with light polarized parallel to the optic axis, no rotation in the polarization of the light occurs in the element's throughput. In this case, the analyzer minimizes transmission through the system, since its transmission axis is perpendicular to both the input polarization and optic axis, see Fig. 1. Applying the opposite voltage to the contacts results in rotating the element's optic axis  $45^\circ$  with respect to its original orientation. The light

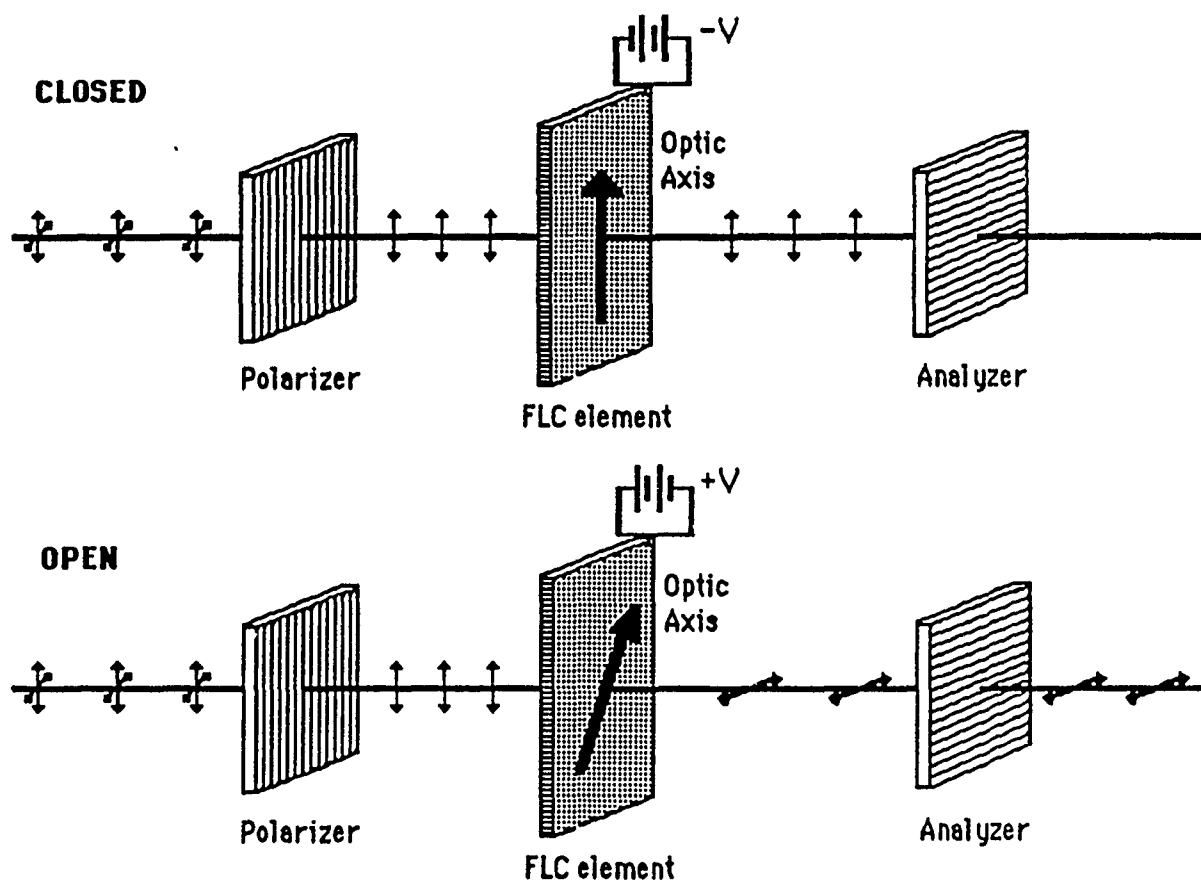


Figure 1. Operating Principles of the the SLM.

transmitted through the array therefore undergoes a  $90^\circ$  rotation in polarization. Consequently, its polarization is parallel to the transmission

axis of the analyzer and results in maximum transmission (see Figure 1, OPEN state). The pattern of light transmitted through the array can thus be regulated by controlling the voltages sent to each element. Further information on the basic principles of the SLM can be found on page 2 of the Operator's Manual.

The interest in investigating a spatial light modulator was to determine the feasibility of its use in controlling the throughput of pixels generated by a binary phase grating element that splits an input laser beam into 25 spatially separated laser beams. For this purpose, a custom made Ferroelectric Liquid Crystal (FLC) spatial light modulator manufactured by Displaytech, Inc. of Boulder, Colorado, was tested to determine its contrast ratios for multiple element and single element areas. This report describes the software and commands used to control the device, the experimental methods implemented to study the device, and the experimental results.

### **3b OPERATING THE SLM**

The device was driven with the Displaytech DDR128 Driver and controlled with its accompanying Pattern Generator board and software. The driver provides the voltages required to run the FLC array whereas the board communicates to the driver the patterns created with the software. Once the board and software were installed in a Zenith 150 computer, any pattern or series of patterns (sequences) could be activated on the FLC array by the user. This section is best understood if the user is operating the computer software as it is read.

Hardware connections and software installation should be performed per the manufacturer's Operator's Manual. The manufacturer warns that leaving the array, or any element, in a particular state for more than 5 seconds could possibly damage its operation. Discussion of other potential operating dangers can be found in the Operator's Manual and the DDR256 Driver for FLC Direct-Drive Arrays User's Manual, Version 1.0.

### **3c PRECAUTIONS**

The manufacturer warns that leaving the array, or any element, in a particular state for more than 5 seconds could possibly damage its operation.

tion. Therefore no single pattern was executed with a step time longer than 4005 msec, though sequences can run repeatedly without time restrictions. In addition, patterns were followed by their inverses for the same durations to insure that no element was operating at hazardous charge build-up levels. The sum of the step times for the two patterns should not exceed 5 seconds as charge build-up harmful to the array could occur in the elements that do not switch in the pattern change.

Further, caution should be exercised when applying optical power to the elements of the array. Displaytech, Inc. specifies that the elements should not be exposed to more than 500 mW/cm<sup>2</sup> (i.e. 5 mW/cm<sup>2</sup>) from a continuous wave laser. For this reason, the optical power density did not exceed 4.5 mW/cm<sup>2</sup> when investigating a single element. Special care should be taken when using focusing optics to illuminate individual pixels.

### **3d EXPERIMENTAL METHOD**

The tests on the SLM were performed at both 830 and 1320 nm diode laser wavelengths. The SLM is intended to operate at a wave length of 1320 nm. The device was characterized by Displaytech, Inc. at 830 nm. Therefore the performance characteristics of the SLM at 830 nm were studied for comparison with those listed by the manufacturer.

As mentioned earlier, the FLC array and output polarizer transmits light when the polarization of the input beam is parallel to an alignment of the liquid crystal molecules. This array was designed so that it would require incident light that was vertically polarized. Thus, the first tests done at each wavelength were to verify the polarization of the input laser beams and to determine the contrast ratio of each polarizer used. A Gala laser served as the 830 nm light source. Using a broadband Melles Griot polarizer and a Newport 835 Optical Power Meter, the Gala laser beam was confirmed to be vertically polarized as the polarizer yielded a contrast ratio of approximately 1000:1. At the 1320 nm wavelength, a "Lightwave" laser was used. It, too was checked for its polarization with the 1320 nm narrow-band polarizer supplied as part of the SLM. This laser also was verified to emit vertically polarized light with the contrast ratio of the polarizer being on the order of 10<sup>5</sup>:1.

An inspection of the SLM and its driver followed the confirmation of the polarization states of each laser. A Newport modular beam-expanding filter consisting of a 15x microscope objective, a 25  $\mu\text{m}$  diameter pin hole, and a collimating lens was inserted into the system to produce a uniformly collimated beam expanded to illuminate the entire array. To image the throughput light, a lens and a screen of facsimile thermal paper were inserted after the SLM, as shown in Figure 2 (not drawn to scale).

A lead-selenide CCD camera with a spectral response from visible wavelengths to 1.6  $\mu\text{m}$  was used to acquire the image information. The generated analog video signal was then fed directly to a TV monitor or to a Macintosh™ II computer. A frame grabber board installed in the computer digitized the image and the software package Image 2.0 allowed for the generation of three-dimensional intensity plots.

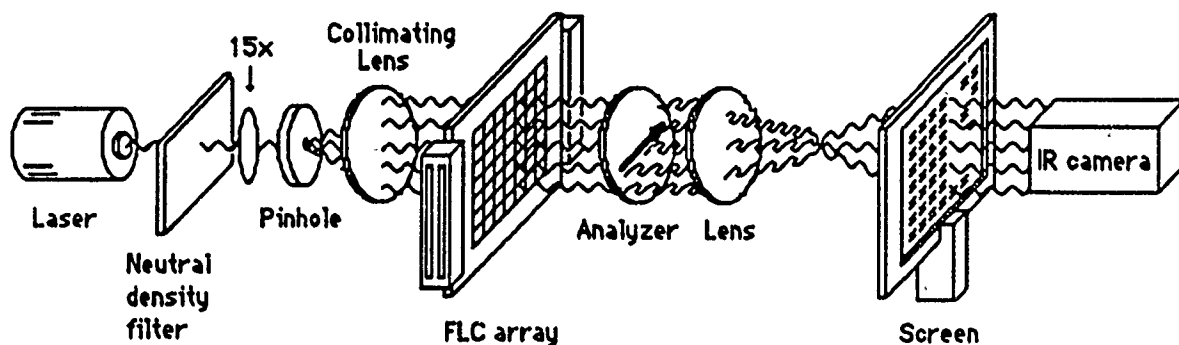


Figure 2. Experimental setup for qualitative observation of the SLM's response to bias.

The experiment revealed that not all of the elements were capable of switching states as designed. A sequence in which all 100 pixels were in the "on" state (i.e. allowing for the passage of light) for 4 seconds and then in the "off" state for 4 seconds, areas of the array unable to respond to the sequencing instructions could be seen. In this device only pixels in half of the array were "on" while only random elements on the other half of the array were "on". Despite programming the entire array to switch on and off, these "dead" pixels failed to respond. Tracing the faulty pixels to their electrode pinouts unveiled that all of the non-operating pixels were controlled through the same channel on one of the connectors. Further in-

vestigation of the channel revealed that port C of the driver failed to output a positive voltage during the sequence. The pins in this port maintained a constant voltage of -14.6 V, whereas the other ports had pins alternating from -14.6 V to 14.6 V. Various pins in these ports did not directly switch between the two voltages; an intermediate voltage was sometimes read before a shift was completed.

Once the functioning status of the FLC array was determined, the setup was modified to test multipixel areas of the array for their contrast ratios. The contrast ratio is the ratio of the light power transmitted when the SLM allows the passage of light (i.e. the "on" state) to that transmitted when in the "off" state (power is at a minimum). Ideally, an input polarizer and an analyzer with mutually perpendicular transmission axes could be rotated synchronously to minimize the off state power through the array. However, since the polarization of the input beam in this setup was supplied by the laser itself, it was necessary to rotate the FLC array and leave the laser and analyzer fixed. A mount was constructed for the array housing such that the device could be rotated about the beam axis.

Because half of the array was not operating correctly, a mask was used to block the light from the defective part of the array so that test efforts would be concentrated on the array's working elements. The optical power meter, connected to a Zenith 248 computer through an IEEE-488 interface, replaced the screen and the camera so that power measurements could be acquired and saved on the hard disk.

To perform a similar study on single elements of the array, the modular beam-expanding filter was removed from the setup. The approximate exit beam dimension of the Lightwave laser was 1.5 mm x 0.7 mm whereas the pixels had dimensions of 0.85 mm x 0.85 mm. Even though each laser was listed as having a beam diameter smaller than the width of an element, a lens was inserted to assure that the diverging laser beam would be focused down to illuminate only the area of a single pixel.

### 3e EXPERIMENTAL RESULTS

A dependence of the contrast ratio on the orientation of the FLC array to the input polarization was evident from the data. Table 2 lists read

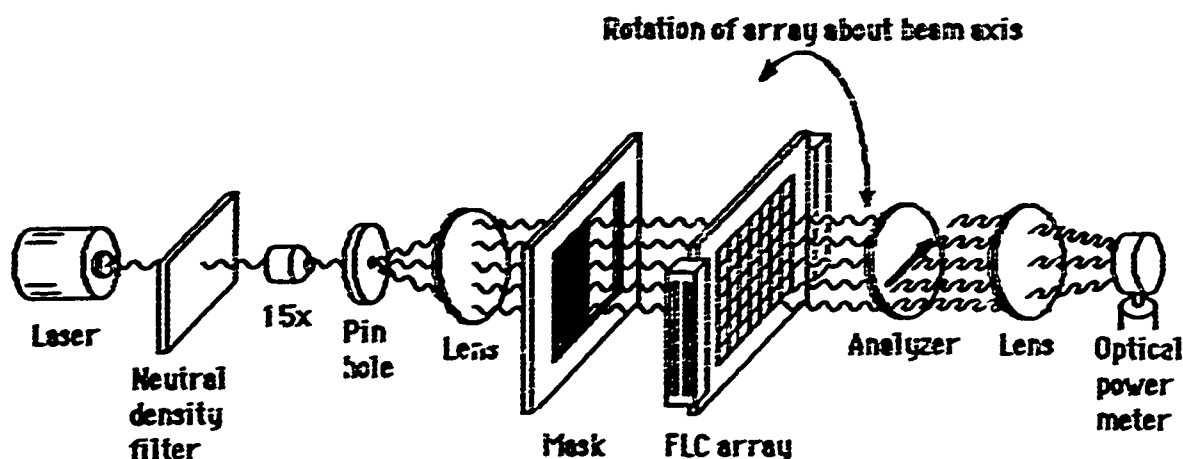


Fig. 3. Experimental setup for power measurements of multiple element areas on the SLM.

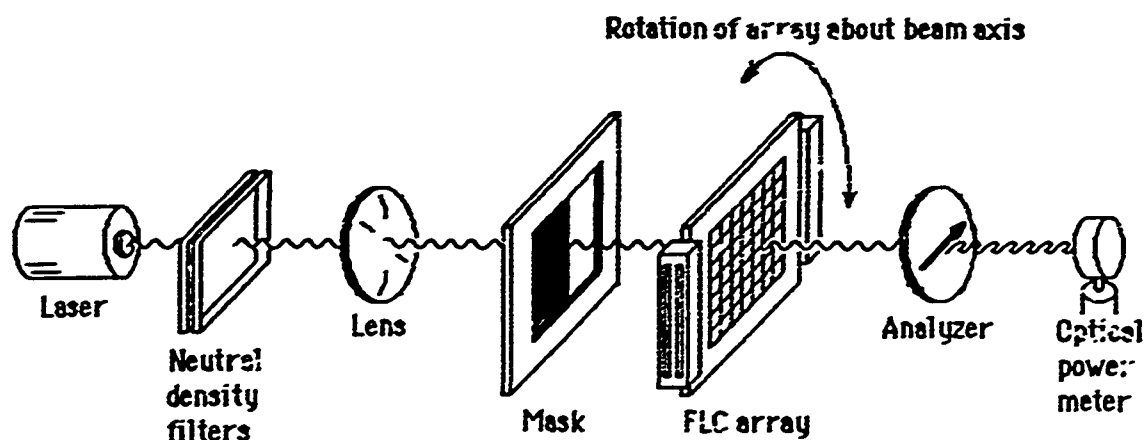


Figure 4. Experimental arrangement for measuring the light transmission of a single pixel.

Rotation of SLM (mount reading)	Power (in $\mu\text{W}$ )		Contrast ratio
	Maximum	Minimum	
$0^\circ$	357	14.6	24.5:1
$3^\circ$	350	7.3	47.9:1
$4^\circ$	344	4.9	70.0:1
$5^\circ$	336	3.3	101.8:1
$5^\circ$	332	3.5	94.0:1
$6^\circ$	318	3.6	88.3:1
$7^\circ$	308	5.0	61.0:1

Table 2 : Test data for multiple element areas of the SLM using the 830 nm source

ings taken to determine the contrast ratio for a multiple pixel area of the SLM using the 830 nm source. The data shows how a change in the contrast ratio was induced by rotating the array. The contrast ratio for the area fell off rapidly from the maximum as the array was rotated through a few degrees of arc.

Further study of this dependence with respect to individual elements revealed that each element had its own orientation for which the contrast ratio was maximized. These orientations varied within a few degrees of the orientation,  $\alpha_{\max}$ , that produced the maximum contrast ratio for the multiple pixel area discussed above. Table 3 displays the readings obtained from four different elements (designated as pixels a, b, c, and d) of the SLM also using 830 nm light. The contrast ratio for each of these elements was greater than that found for the multiple pixel area, but was in agreement with Displaytech, Inc.'s measured contrast ratio of 203:1 for a single element. The weaker performance of the larger area directly stemmed from the requirement that each element needed a different orientation to maximize its contrast ratio.

The samples taken while using the 830 nm source also suggested that the two states of the optic axis for the multiple element area did not differ by 45 degrees. If the two orientations were at a 45 degree

	Rotation of SLM (mount reading)	Power (in $\mu\text{W}$ )		Contrast ratio
		Maximum	Minimum	
Pixel a.	$3^\circ$	905	8.1	112:1
	$5.5^\circ$	479	3.9	123:1
	$5.5^\circ$	453	3.8	119:1
Pixel b.	$3.5^\circ$	759	4.5	169:1
	$4.5^\circ$	900	4.4	205:1
	$4.5^\circ$	937	3.8	247:1
Pixel c.	$3.5^\circ$	1201	5.4	222:1
	$3.5^\circ$	1105	5.1	217:1
Pixel d.	$4.5^\circ$	786	4.0	197:1
	$4.5^\circ$	945	4.4	215:1

Table 3: Test data for single elements of the SLM at 830 nm



<u>Rotation of SLM (mount reading)</u>		<u>Power (in <math>\mu\text{W}</math>)</u>		<u>Contrast ratio</u>
		<u>Maximum</u>	<u>Minimum</u>	
Pixel e.	3 <sup>0</sup>	494	21.9	23:1
Pixel f.	9.5 <sup>0</sup>	1003	3.2	313:1
	9.5 <sup>0</sup>	1079	3.3	327:1
Pixel g.	9.5 <sup>0</sup>	1370	2.3	596:1
Pixel h.	10 <sup>0</sup>	1354	0.54	2571:1

Table 4: Test data for single elements of the SLM at 1320 nm

<u>Pixel index (rotation fixed at 9.5<sup>0</sup>)</u>	<u>Power (in <math>\mu\text{W}</math>)</u>		<u>Contrast ratio</u>
	<u>Maximum</u>	<u>Minimum</u>	
1	1360	3	450:1
2	1450	6	240:1
3	1330	1	1330:1
4	1380	2	690:1
5	1430	5.5	260:1
6	1390	4	350:1
7	1410	1.5	940:1
8	1400	1	1400:1

Table 5: Test data for random elements of the SLM at 1320 nm  
(rotation fixed)

separation, similar power readings would be seen at angles  $+\theta$  and  $-\theta$  away from  $q_{\text{max}}$ . The maximum power would decrease as the array was rotated away from  $q_{\text{max}}$ , whereas the minimum power would increase. These tendencies were not present in the measurements. The maximum power column in Table 2 reflects the non-symmetry in power readings that suggests the two states of the multiple element area were not 45 degrees apart. The output power was seen to increase when the array was rotated away from  $q_{\text{max}}$  in one direction, but to decrease when done so in the opposite direction. The separation between the two states may not remain constant as the area is rotated. Further investigation could determine the separation between the two states of a multiple element area and its dependence on the orientation of the array to the analyzer.

The measurements completed using the 1320 nm source resulted in significantly higher contrast ratios than those taken using 830 nm light.

Table 4 lists measurements of the power outputs for four random elements to determine the contrast ratio that could be expected for an individual element illuminated with 1320 nm light. The first reading may not have been rotated to yield the best contrast ratio, whereas the others were taken at the orientation giving the lowest "off" state power transmission. For comparison, a contrast ratio for a multiple element area was measured to be 107:1.

Readings taken of eight random elements with the rotation of the array fixed to one orientation also demonstrated how each element had to be reoriented to maximize its contrast ratio; see Table 5. These measurements further suggested that the SLM functioned better when illuminated with 1320 nm light as the contrast ratios were greater than those made using the 830 nm source.

It is important to note that the magnitude of the contrast ratio is limited by the quality of the polarizers used in the polarizer-analyzer system. If the system has a maximum to minimum transmission ratio of 250:1, then the SLM cannot have a contrast ratio greater than 250:1. Although the SLM appeared to function better at 1320 nm than at 830 nm, the analyzer at 1320 nm extinguished light with greater efficiency than the 830 nm polarizer. The difference in performance of these polarizers at their respective wavelengths may have accounted for the significant improvement in the contrast ratio of the device when studied using 1320 nm light.

Error may have been introduced as a result of the large background noise present during power measurements. The background noise varied from 4  $\mu\text{W}$  to 14  $\mu\text{W}$  with a fluctuation of approximately 0.5  $\mu\text{W}$ . This noise was 1 to 4 times greater than the output power to be measured when the SLM was operating in the "off" state. The background was subtracted from the readings using one of two processes. When measurements were acquired through the IEEE-488 interface, the background was calculated by averaging 50 samples of the power meter readout with the source beam blocked before a series of measurements was conducted. This average background was subtracted from each measurement before it was stored for future data processing. For measurements obtained by a visual reading of the power meter's LCD display, a null button was activated while the source beam was blocked. The null button allowed for the storage of the

reading present on the display when the button was depressed. The power meter would then subtract the stored value from any future readings before they were displayed.

The overall assessment of the SLM is that it is not optimal for the intended application where it would be used to independently switch 25 spatially separated laser beams. The switching requirement would easily be satisfied by the manufacturer's design of the controller, driver, and software, but the need to rotate the array to maximize the throughput for each element disqualifies the array as a candidate for the modulation of the low power laser beams that will be present in the system. The variation in contrast ratio among individual pixels at a fixed angle (see Table 5) indicates that the throughput of some elements would not be maximized without rotation of the array. The SLM is generally used with a laser beam of sufficient power so that the reduced throughput would not adversely affect the dynamic range of a detection system. In this application, the decreased throughput that will result in the pixels that are not oriented at their maximum would cause detection of the low power laser beams to be close to the noise level of a photodetector.

The splitting of the laser beam into symmetric, low power beams is a basic design criteria and cannot be changed. Therefore, another scheme must be employed to switch the 25 spatially separated laser beams.

## 4 BINARY LENS

We tested a binary lens. This device replicates a single light beam in a 25 by 25 beam array. The binary lense consists of a glass plate in which a two dimensional grading like pattern was etched. The pattern of this particular device has only two levels, the top surface of the glass plate and the etched level. The etched area is about 6 by 6 mm. The lens was designed to operate at a wavelength of 1.3  $\mu\text{m}$ . We tested the device both with a uniform collimated beam from a spatially filtered laser beam and a light beam directly from a laser. The device had an uniformity from beam to beam of 60%. The beam diverges when it emerges from the lense. It has to be recollimated with a lense.

The devise was, first, tested by projecting the light output of the lense on a screen. A lead-selenide CCD camera with a spectral response from

visible wavelengths to  $1.6\text{ }\mu\text{m}$  was used to acquire the image information from the screen. The analog video signal generated was then fed directly to a TV monitor or to a Macintosh™ II computer. A frame grabber board installed in the computer digitized the image and the software package Image 2.0 allowed for the generation of three-dimensional intensity plots.

The device was also tested by measuring the light power in each output beam with an optical power meter. The uniformity of the device was determined with the power meter.

## CHARACTERIZATION OF RADAR CLUTTER AS AN SIRP

CHARLES T. WIDENER and JAY K. LEE, Ph.D.

### ABSTRACT

It has been proposed that radar clutter can be modeled as a spherically invariant random process or SIRP. SIRPs seem well suited to this role since by variation of certain parameters the Weibull, K- or Rayleigh distributed clutter envelopes are obtained. These distributions are significant since they fit well with experimental radar clutter data under different circumstances. In this report, a radar clutter model based on rough surface scattering is developed to show that SIRP characterization can be based on electromagnetic principles. Small perturbation analysis of a two-scale randomly rough surface is chosen since the form for the backscattered field has the proper form for an SIRP, under certain conditions. This is an important step in being able to predict the proper statistical distribution of radar clutter based on surface geometry and electromagnetic properties.

### INTRODUCTION

The characterization of radar clutter has been an important study for radar engineers, from the inception of radar, to the present. Due to the random nature of background terrain, statistical characterization seems appropriate from both theoretical and experimental considerations. Gaussian models have been commonly used, but they become inadequate in many cases of interest. Other characterizations have been proposed on the basis of empirical studies, examples being Rician [1], Weibull [2], log-normal [3], and K-type [4, 5]. More recently it has been proposed that radar clutter be

modeled as a complex random process [6], more specifically as a spherically invariant random process (SIRP) [7-9]. Briefly, an SIRP is such that given a real or complex random process

$$\bar{X}(t) = \{X_1(t), X_2(t), \dots, X_n(t)\}^T$$

every sample taken from it, that is

$$\bar{X} = \{X_1(t_1), X_2(t_2), \dots, X_n(t_n)\}^T$$

is spherically invariant random vector (SIRV) with one and the same characteristic cumulative distribution function (CCDF). An SIRV can be thought of as being obtained by a product of a Gaussian random vector times an independent non-negative random scale factor  $s$  with some CDF, which is the CCDF of SIRV under consideration. The usefulness of the SIRP characterization comes by varying parameters of  $s$  to obtain different clutter statistics. It can be shown [6] that for certain parameters the Rayleigh, K-, or Weibull clutter envelopes result. This unifying concept is an important step in further understanding clutter.

The purpose of this report is to develop a mathematical model for clutter from electromagnetic theory, suitable as a basis for an SIRP characterization. Specifically, the form of received signal (from clutter) should be of the form  $V(t) = G(t) \cdot s$  where  $G(t)$  is a Gaussian random process,  $s$  is a random variable, and  $V(t)$  is the received signal. Without regard to type of receiver, the signal  $V(t)$  sought is the backscattered field  $\vec{E}^s(t)$ . As a further specialization, concentration will be on a single sample (or instance of time) corresponding to the returns from a single radar illumination area or "footprint".

#### DISCUSSION OF APPROPRIATE MODELS

In the attempt to derive useful, accurate mathematical models for radar clutter,

two approaches are commonly used. The first views the illuminated area or target background as a collection of point scatterers [10]. These scatterers are assumed randomly distributed over a surface, each having an associated amplitude and phase. In the most general case, the amplitude and phase distributions are random variables. Classically, by making various assumptions about these distributions, different clutter characteristics are derived, e.g. a Rayleigh distributed envelope results when the phase is assumed equal and constant. This phenomenological approach is useful, but not really derived from first principles.

A second approach, and the one used in this report, is treating radar clutter as scattering from a rough, randomly distributed surface. A rough surface can be described in a system of Cartesian coordinates as

$$Z = \zeta(x, y)$$

where  $\zeta$  is a random variable depending on  $x$  and  $y$ . The mean level of the surface is the plane

$$Z = 0$$

Any point on the surface can be described by a position vector from the origin

$$\vec{r} = \hat{x}x + \hat{y}y + \hat{z}\zeta(x, y)$$

A random surface can be described by the statistical distribution of its deviation from a mean level. This, however, does not completely describe the surface; it says nothing about how close or far apart the hills and depressions are. A second function, the correlation function, or its normalized version, the correlation coefficient describes this aspect of the surface. The correlation coefficient  $C(\tau)$  gives a measure of dependence or

correlation of two points  $\zeta(x_1, y_1)$  and  $\zeta(x_2, y_2)$  on the surface. If the separation of the points

$$\tau = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

is very large, (not on the same hill or valley) then the points are independent. Conversely, points near one another will be correlated; when  $\tau=0$ , they are the same point, or

$$\lim_{\tau \rightarrow 0} C(\tau) = 1$$

The distance  $T$  in which  $C(\tau)$  drops to  $e^{-1}$  is called the correlation length.

A commonly used distribution is the normal distribution with zero mean

$$w(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{z^2}{2\sigma^2}\right)$$

where  $\sigma$  is the rms value of the deviation of  $\zeta$  from  $z=0$ . A commonly used correlation coefficient is

$$C(\tau) = \exp\left(-\tau^2/T^2\right)$$

Specification of  $\sigma$  and  $T$  makes the rough surface model approximate a wide variety of rough surfaces met in practice.

Scattering of waves (acoustic or electromagnetic) from a rough surface is a complex problem studied by many people in this century and yet it remains a popular topic for research papers. Many methods of analysis have been used to study scattering from a rough surface, but usually each method assumes a priori some condition or scale limitation to make the problem tractable.

Two important examples of the methods are the Kirchhoff approximation (or tangent plane approximation) and the small perturbation method (SPM). In the former an assumption is made that the incident wavelength is much smaller than the variations of the surface. Under this assumption, an



incident wave locally "sees" a planar surface. Reflection (and transmission) at a planar interface is a classic problem in electromagnetics, so by integrating over the surface, a solution can be found. Mathematically, the validity of the solution is restricted by the condition, derived by Brekhovskikh [11], that the radius of curvature of the surface is much greater than the incident wavelength or

$$4\pi r_c \cos \vartheta \gg \lambda \quad (\text{as reported by Beckmann})$$

where  $r_c$  is the minimum radius of curvature in any direction and  $\vartheta$  is the angle between the incident propagation vector  $\vec{k}$  and the local normal  $\vec{n}$ . A second condition [10], which intuitively needs to be satisfied, given the condition on  $r_c$  is; the correlation length must exceed the wavelength

$$T > \lambda$$

(Beckmann's formulation [see 10] is classic and widely used.)

Sometimes a condition is used for stationary-phase or geometric optics, which includes that of physical optics, [12, 13] namely

$$4k^2 \cos^2 \theta \sigma^2 \gg 1$$

This has the advantage that the rms height deviation  $\sigma$  is related directly to  $\lambda$  ( $k = \frac{2\pi}{\lambda}$ ). Because the rms height  $\sigma$  and correlation length  $T$  are large compared to  $\lambda$ , this type of surface is called large scale rough surface.

At the other extreme is the SPM. In this method it is assumed that the rms height is small compared to the wavelength

$$4k^2 \cos^2 \theta \sigma^2 \ll 1$$

where  $\theta$  is the angle of incidence.

Rice solved this problem using a two-dimensional Fourier series expansion of the surface [14]. The coefficients of the Fourier components are deter-

mined from boundary conditions on the surface. A variation on this method was developed by Burrows [15], where by also assuming the rms slope is small,

$$(\nabla_s \zeta)^2 \ll 1$$

the surface equation can be expanded in a Taylor series about its mean. Using the small slope assumption, higher order terms can be neglected. Then, in a manner similar to Rice, the exact boundary conditions on the perturbed surface are rewritten as conditions which apply to the unperturbed planar surface and a solution is found. As may be guessed such surfaces are called small scale.

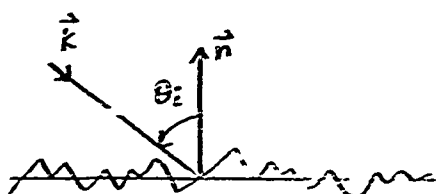
When modeling surfaces occurring in nature, the assumption of exclusively large or small scale is restrictive and unrealistic when compared to modern radar wavelength. If, for example, an ocean were to be modeled as a large scale surface, the radar wavelength would need to be submillimeter due to the presence of fine, wind produced ripples. Conversely, if it were modeled as a small scale surface, the corresponding radar wavelength would be on the order of tens of meters, because of the large sea swells or gravity waves. Such is the case with most surfaces of interest with regard to radar clutter.

To accommodate analysis of more realistic surfaces, a two-scale or composite model has been proposed and studied by several authors [12, 16-21]. A composite surface is simply viewed as a superposition of a small scale variation on a large scale variation.

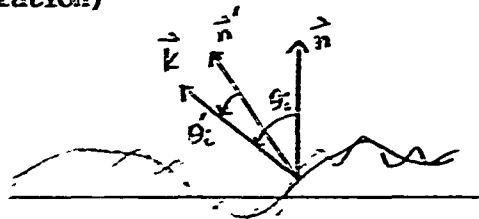
$$\zeta = \zeta_L + \zeta_S$$

The division between large scale and small scale occurs naturally once the radar wavelength is specified. While this model has some limitations of its own, it is far more versatile than either the large or small scale model alone.

Two of the more interesting and simple approaches to dealing with composite surfaces are the tilted plane method and a more generalized SPM. In the tilted plane method, the large scale variation is assumed to present locally flat titled plane areas with superimposed small scale variation. The only difference in the analysis between this and any other small scale model is that the angle of incidence with regard to the normal has effectively undergone a spatial rotation. (see illustration)



a) wave incident upon small scale surface



b) wave incident upon composite surface

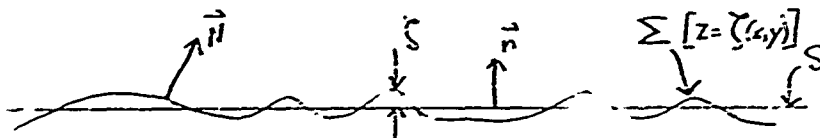
The SPM applied to a composite surface is also similar to that applied to an exclusively small scale surface. The exact boundary conditions on the surface are transferred to a system of boundary conditions on the large scale surface instead of the surface  $z=0$ . Effectively, the variations of scattered field due to the small scale roughness are viewed as being caused by equivalent surface current densities which create the same resultant field as the small scale roughness. The total scattered field is the sum of zeroth order component, due only to the small scale variation. Brown [12] and Bass et al. [19, 20] have both applied this method to a composite scale surface.

Because of the relatively simple form of the scattered field coupled with the fact that cross polarization terms are built in the final form, the SPM for composite surface is useful as a basis to characterize radar clutter as an SIRP. A short summary of this method appears in the following section.

#### SUMMARY OF SMALL PERTURBATION ANALYSIS

The first case treated is that of a small scale surface only. This is done to illustrate the main points of the analysis without becoming unduly bogged down in cumbersome details. Following this summary, it is shown how small perturbation analysis is applied to a composite surface as a logical next step.

In the scattering analysis of a small scale surface, the surface (denoted  $\Sigma$ ) is represented by normal deviation  $\zeta(\vec{r})$  from a smooth surface  $z=0$  denoted  $S$ .



The normal to the surface  $\Sigma$  is denoted  $\vec{N}$ , while the normal to the surface  $z=0$  is denoted  $\vec{n}$ . It is assumed that the rough surface divides a vacuum and medium with relative permittivity  $\epsilon$ . Also it is assumed that the permeability  $\mu$  is constant across the interface. Boundary conditions on the surface  $\Sigma$  have the form

$$\begin{aligned} [\vec{N} \times (\vec{E}_1 - \vec{E}_2)]_{\Sigma} &= 0, \quad [\vec{N} \cdot (\vec{E}_1 - \epsilon \vec{E}_2)]_{\Sigma} = 0 \\ [\vec{H}_1 - \vec{H}_2]_{\Sigma} &= 0 \end{aligned} \quad (1)$$

The subscript 1 refers to the fields in the first medium, while the subscript 2 refers to the fields in the second medium. Assuming the normal

deviations  $\zeta$  and the slopes  $\vec{\gamma} = \nabla \zeta$  of the surface  $\Sigma$  relative to the surface  $S_0$  are small

$$(k\sigma)^2 \ll 1, \quad \vec{\gamma}^2 \ll 1 \quad (2)$$

where  $\nabla_S$  is a surface gradient and  $\sigma$  is the rms deviation of  $\zeta$ , the boundary condition (1) may be transferred to the mean surface  $S$  through a Taylor series expansion. Higher order terms can be neglected by the conditions (2). We assume the fields are of the form of a mean field  $\vec{E}$  and a fluctuation field  $\vec{e}$ .

$$\vec{E} = \vec{E} + \vec{e}, \quad \vec{H} = \vec{H} + \vec{h} \quad (3)$$

In this approximation, the mean field coincides with the field reflected from  $S$ , while the fluctuation field is small relative to the mean. The boundary conditions are thus rewritten

$$[\vec{n} \times (\vec{E}_1 - \vec{E}_2)]_S = 0, \quad [\vec{n} \cdot (\vec{E}_1 - \epsilon \vec{E}_2)]_S = 0, \quad \vec{H}_1|_S = \vec{H}_2|_S \quad (4)$$

$$\left. \begin{aligned} \vec{n} \times (\vec{e}_1 - \vec{e}_2) &= [\vec{\gamma} \times (\vec{E}_1 - \vec{E}_2)] - [\vec{n} \times \frac{\partial}{\partial z} (\vec{E}_1 - \vec{E}_2)] \zeta \\ \vec{n} \times (\vec{h}_1 - \vec{h}_2) &= [\vec{\gamma} \times (\vec{H}_1 - \vec{H}_2)] - [\vec{n} \times \frac{\partial}{\partial z} (\vec{H}_1 - \vec{H}_2)] \zeta \end{aligned} \right\} \quad (5)$$

Equivalent current densities  $\vec{j}_e$  and  $\vec{j}_m$ , electric and magnetic, respectively, are defined as

$$\vec{n} \times (\vec{e}_1 - \vec{e}_2) = -\frac{4\pi}{c} \vec{j}_m, \quad \vec{n} \times (\vec{h}_1 - \vec{h}_2) = \frac{4\pi}{c} \vec{j}_e \quad (6)$$

by expressing  $\vec{\epsilon}_2$  in terms of  $\vec{\epsilon}_1$  (through boundary conditions 4) and  $\vec{H}$  in terms of  $\vec{E}$  (through Maxwell's eqs)  $\vec{j}_m$  and  $\vec{j}_e$  are expressed in terms of  $\vec{E}$ ,  $\vec{n}$ ,  $\zeta$  as

$$\begin{aligned}\vec{j}_e &= i k c \frac{1-\epsilon}{4\pi} [\vec{n} \times (\vec{E} \times \vec{n})] \zeta \\ \vec{j}_m &= \frac{c}{4\pi} \frac{\epsilon-1}{\epsilon} [\vec{n} \times \nabla(\vec{n} \cdot \vec{E}) \zeta]\end{aligned}\quad (7)$$

The general field  $\vec{E}_{(2)}$  at a point above (below) the dielectric is expressed as a sum of plane waves, i.e.

$$\vec{E}_{1,2}(\vec{R}) = \int_{-\infty}^{\infty} \tilde{E}_{1,2}(\vec{k}) \exp[i(\vec{k} \cdot \vec{r}) + \sqrt{k_{1,2}^2 - k^2} z] d\vec{k} \quad (8)$$

where  $\vec{R} = \vec{r} + \hat{z} z$ ,  $\vec{k} = |k| \vec{\beta}$ ,  $\vec{\beta} = \vec{\beta}_\perp + \hat{z} \beta_z = \frac{\vec{k}}{|\vec{k}|}$

Each component plane wave must satisfy (from Maxwell's eqs)

$$\begin{aligned}\vec{h}_1 \times \vec{\beta} &= \tilde{e}_1, \quad [\vec{n} \times (\vec{h}_1 - \vec{h}_2)] = \frac{4\pi}{c} \tilde{j}_e \\ \vec{e}_1 \times \vec{\beta} &= -\vec{h}_1, \quad [\vec{n} \times (\vec{e}_1 - \vec{e}_2)] = -\frac{4\pi}{c} \tilde{j}_m \\ \vec{h}_2 \times (\vec{\beta} - (a+b)\vec{n}) &= \epsilon \tilde{e}_2, \quad [\tilde{e}_2 \times (\vec{\beta} - (a+b)\vec{n})] = -\vec{h}_2\end{aligned}\quad (9)$$

where

$$\tilde{j}_{e,m}(\vec{k}) = \frac{1}{(2\pi)^2} \int \vec{j}_{e,m} e^{-i\vec{k} \cdot \vec{r}} d\vec{r} \quad (10)$$

and

$$a = \vec{n} \cdot \vec{\beta}, \quad b = \sqrt{\epsilon - 1 + a^2}$$

From these expressions,  $\tilde{e}_1(\vec{k})$ ,  $\tilde{e}_2(\vec{k})$  are found to be

$$\tilde{e}_1(\vec{k}) = \frac{4\pi}{c(b+a\epsilon)} \left\{ \frac{\epsilon-1}{a+b} (\vec{\beta} \cdot \tilde{j}_m) (\vec{n} \times \vec{\beta}) + [\vec{\beta} \times (\vec{n} \times \tilde{j}_e)] + \epsilon (\vec{\beta} \times \tilde{j}_m) + [\vec{\beta} \times (\vec{\beta} \times \tilde{j}_e)] \right\} \quad (11)$$

$$\tilde{e}_2 = \tilde{e}_1 + \frac{1-\epsilon}{\epsilon} \vec{n} \cdot (\vec{n} \cdot \tilde{e}_1) + \frac{4\pi}{c} (\tilde{j}_m \times \vec{n}) + \frac{4\pi}{c\epsilon} \vec{n} (\vec{\beta} \cdot \tilde{j}_e)$$

The field  $\tilde{E}_1$ , on the surface  $z=0$  is written using the Fresnel formulas for an incident field  $\vec{E}^0$

$$\text{where } \vec{E}^0 = \hat{p} e^{ikR_0} e^{i(\vec{k}_1 \cdot \vec{r} + k_z z)} \quad (12)$$

where  $R_0$  is the distance from the source to the center of the scattering surface and  $\hat{p}$  is the polarization vector. Thus it can be derived

$$\vec{E}_1 = \frac{2a_0}{a_0+b_0} \vec{E}^0 - \frac{2a_0(1-\epsilon)}{b_0+\epsilon a_0} \left( \vec{n} + \frac{\vec{\alpha}}{a_0+b_0} \right) (\vec{E}^0 \cdot \vec{n}) \quad (13)$$

where

$$a_0 = -\vec{n} \cdot \vec{\alpha}$$

$$b_0 = \sqrt{\epsilon - \cos^2 \psi}$$

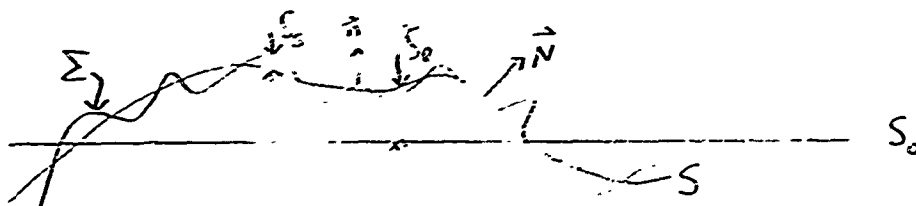


Substituting eq (13) for  $\vec{E}_1$  into eqs (7) for  $\tilde{j}_e$  and  $\tilde{j}_m$ ; eqs (7) into eqs (10) for  $\tilde{j}_e$  and  $\tilde{j}_m$ ; and eqs (10) into eq (8), the field is found at any point in space. The integral may be evaluated by saddle-point integration in view of the fact that the observation point is located in the Fraunhofer zone. For the backscatter case where  $\vec{k}_i = -\vec{k}_r$ ,

$$\vec{E} \cdot \hat{p} = \frac{k^2(\epsilon-1)}{\pi R_0} e^{2ikR_0} E^0 \int_A \left\{ \left( \frac{a}{a+b} \right)^2 (\hat{p} \cdot \hat{p}_0) + \frac{2(\epsilon-1)a^2b}{(b+a\epsilon)^2(a+b)} \right. \\ \left. \cdot (\vec{n} \times \hat{p}) (\vec{n} \times \hat{p}_0) \right\} e^{2ik\vec{\beta} \cdot \vec{r}} \zeta(\vec{r}) d\vec{r} \quad (14)$$

For the case of a small scale  $\lambda$  <sup>surface</sup> only the zeroth order field  $\vec{E}$  is simply the

and can be field reflected from a plane interface, found using Fresnel coefficients. When the SPM is applied to a composite surface, the methodology is similar, but more cumbersome. For the two scale case, we have a surface, denoted by  $\Sigma$ , represented by normal deviations  $\zeta(\vec{r}_s)$  from a smoother surface  $S$ .



The position of points on  $\Sigma$  is related to points on  $S$  by  $\vec{r}_\Sigma = \vec{r}_s + \vec{N}\zeta(\vec{r}_s)$ , where  $\vec{N}$  is the normal to the mean surface  $S$ . The normal to  $\Sigma$  will be denoted  $\vec{n}$ . For a medium of relative permittivity  $\epsilon$  below the surface and vacuum above, the boundary conditions are

$$\begin{aligned} [\vec{n} \times (\vec{E}_1 - \vec{E}_2)]_\Sigma &= 0, \quad [\vec{n} \cdot (\vec{E}_1 - \epsilon \vec{E}_2)]_\Sigma = 0 \\ (\vec{H}_1 - \vec{H}_2)|_\Sigma &= 0 \end{aligned} \quad (15)$$

Again, assuming the surface deviations from  $S$  are small, with gentle slopes such that

$$(k\sigma)^2 \ll 1, \quad (\nabla_s \zeta)^2 \ll 1 \quad (16)$$

where  $\sigma^2 = \text{Var}[\zeta]$  relative to  $S$ , then an expansion of the fields in terms of powers of  $\zeta$  is valid, keeping only the first two terms

$$\begin{aligned} \vec{E}(\vec{r}_\Sigma) &\approx \vec{E}(\vec{r}_s) + \zeta(\vec{r}_s)(\vec{N} \cdot \nabla)\vec{E}(\vec{r}_s), \quad \vec{n} \approx \vec{N} - \vec{\gamma} \\ \vec{H}(\vec{r}_\Sigma) &\approx \vec{H}(\vec{r}_s) + \zeta(\vec{r}_s)(\vec{N} \cdot \nabla)\vec{H}(\vec{r}_s), \quad \vec{\gamma} = \nabla_s \zeta(\vec{r}_s) \end{aligned} \quad (17)$$



The scattered field is again assumed to be of the form  $\vec{E} = \vec{E}_0 + \vec{e}$ ,  $\vec{H} = \vec{H}_0 + \vec{h}$  and the analysis proceeds in a similar fashion to the analysis for a small scale surface only. In both cases the final form for the field scattered from the small scale roughness is the same, the only difference being  $k_z \zeta_0$  appears in the exponential  $e^{-j2\vec{k} \cdot \vec{R}}$  for the composite surface case. If the limiting case is taken,  $\lim_{\epsilon \rightarrow \infty}$  (perfectly conducting surface), then the field becomes

$$\vec{e} = \frac{E^0 k^2}{\pi R_0} e^{-jkR_0} \iint \frac{2(\vec{n} \cdot \hat{p})(\vec{n} \cdot \hat{p}_0) + (\vec{n} \cdot \vec{k}_i)^2 (\hat{p} \cdot \hat{p}_0)}{\sqrt{1 + (\nabla_s \zeta_0)^2}^{-1}} (18)$$

$$- \zeta_s e^{-j2(\vec{k}_i \cdot \vec{r} + \zeta_0 k_z)} dx dy$$

a simpler form of (14) to work with. The zeroth order field is found in a much more direct fashion to be [12]

$$\vec{E}_0 = \frac{-jkE^0 \delta_{pp'}}{2\pi R_0 \cos \theta} e^{-jkR_0} \iint e^{-j2(\vec{k} \cdot \vec{r} + k_z \zeta_0)} dx dy \quad (19)$$

for the perfectly conducting surface case.

#### DEVELOPMENT OF THE SIRV FORM

In general the SPM applied to composite surface of large and small scale variations,  $\zeta_0$  and  $\zeta_s$ , respectively, predicts a scattered field

$$\vec{E}^s = \vec{E}_0 + \vec{e}$$

i.e. a sum of a zeroth order field and a first order fluctuation field.

Brown derives these components (for a perfectly conducting surface) as:

$$E_{PP'} = \frac{E^0 k^2}{\pi r} \exp(-jk r) \iint_A \frac{2(\vec{n} \cdot \hat{p})(\vec{n} \cdot \hat{p}') + (\vec{n} \cdot \vec{k}_i)^2 (\hat{p} \cdot \hat{p}')}{\sqrt{1 + \zeta_{2x}^2 + \zeta_{2y}^2 - 1}} \cdot \zeta_e \exp(-j 2 \vec{k}_i \cdot \vec{r}_0) dx dy$$

$$E_{0PP'} = \frac{-j E^0 k \delta_{PP'}}{2 \pi r \cos \theta} \exp(-jk r) \iint_A \exp(-j 2 \vec{k}_i \cdot \vec{r}_0) dx dy$$

where  $E^0$  is the amplitude of incident field

$\vec{k}_i$  specifies the propagation vector of the incident field

$\vec{r}_0 = \hat{x}x + \hat{y}y + \hat{z}\zeta_e$  is the position vector from the origin to any point on the large scale surface

$\vec{n}$  is the unit normal to the large scale surface

$k = |\vec{k}_i|$  is the wavenumber

$\hat{p}$ ,  $\hat{p}'$  denote incident and backscattered polarizations, respectively

$$\zeta_{ex} = \frac{\partial \zeta_e}{\partial x}, \quad \zeta_{ey} = \frac{\partial \zeta_e}{\partial y}$$

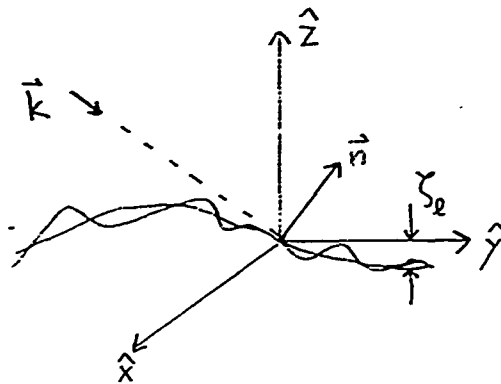
$\hat{k}_i = \frac{\vec{k}_i}{k}$  a unit vector in the direction of incident propagation

$r$  is the distance from the transmitter-receiver to the origin

$$\delta_{PP'} = \begin{cases} 1 & \text{if } \hat{p} = \hat{p}' \\ 0 & \text{if } \hat{p} \neq \hat{p}' \end{cases}$$

$A$  represents the beam illuminated area (radar footprint)

$\theta$  is the angle of incidence =  $\angle(-\vec{k}_i, \hat{z})$



The incident field  $\vec{E}_i = \hat{p} E^0 e^{-j \vec{k}_i \cdot \vec{r}}$  is assumed to be a plane wave with time dependence  $e^{j\omega t}$  suppressed.

As previously mentioned the term  $\vec{E}_0$  represents the scattered field from the large scale surface variation only. It is well known that the  $\vec{E}_0$  term is strongest in the specular direction and falls off rapidly outside a cone centered axially on the specular wavevector. When this general result is specialized to radar backscatter the implication is that the zeroth order field has its strongest effect on the return signal at or near normal incidence  $\theta = 0^\circ$ . It will in fact dominate the radar return if the radar receiver is located within the cone. It can also be shown [10, chap 5] that the spread of the cone is proportional to rms height cone width (in degrees)  $\propto \sigma/\lambda$ . In the limiting case as  $\sigma \rightarrow 0$  the surface becomes a plane and reflection occurs in the specular direction only (Snell's Law), which is intuitively satisfying.

Now the desired mathematical form sought is a product of a Gaussian random variable and some scalar. To simplify the form of the backscattered field we assume that the slope  $m$  of the large scale variation is less than 0.2. Ulaby, Moore and Fung [13, chap 12] have calculated backscattering coefficients due to large scale roughness using the tangent plane method which shows

that large scale backscatter becomes very small for angles of incidence greater than  $30^\circ$  if the large scale rms slope is less than 0.2. Under these conditions, the first order field  $\vec{E}$  will dominate the radar backscatter. By grouping terms of  $\vec{E}$  which depend only on  $\zeta_\ell$ ,  $\zeta_s$ , or  $\nabla\zeta_\ell$ , the scattered field  $\vec{E}^s$  has the form (subject to the aforementioned limitation)

$$\vec{E}^s \approx \vec{E} = C \iint_A \Gamma_{pp'}(\nabla\zeta_\ell) \zeta_s e^{-j2(\vec{k} \cdot \vec{r}_{xy} + k_z \zeta_\ell)} dx dy$$

where  $C = \frac{E^0 k^2}{\pi r} e^{-jk r}$

$$\Gamma_{pp'}(\nabla\zeta_\ell) = [2(\vec{n} \cdot \hat{p})(\vec{n} \cdot \hat{p}') + (\vec{n} \cdot \vec{k}_i)^2 (\hat{p} \cdot \hat{p}')] \sqrt{1 + \zeta_{\ell x}^2 + \zeta_{\ell y}^2}$$

The basic area of integration A for a single look of the radar (sometimes called the radar footprint) can be broken up into a summation over smaller subareas or patches. Since  $\zeta_\ell$  and  $\nabla\zeta_\ell$  vary much more "slowly" over the footprint, a logical way to break up the integral is by patches over which  $\zeta_\ell$  and  $\nabla\zeta_\ell$  remain fairly constant. Doing so yields

$$\vec{E}^s = C \sum_p \Gamma[\zeta_\ell(p)] e^{-j2k_z \zeta_\ell(p)} \iint_{\Delta p} \zeta_s(x,y) e^{-j2(\vec{k} \cdot \vec{r}_{xy})} dx dy$$

or more compactly

$$\vec{E}^s = \sum_p Z(p) \iint_{\Delta p} \zeta_s(x,y) e^{-j2(\vec{k} \cdot \vec{r}_{xy})} dx dy$$

where

$$Z(p) = C \Gamma[\zeta_\ell(p)]$$

$\vec{E}^s$  is now of the form

$$\vec{E}^s = Z_1 \iint_{\Delta 1} [ \quad ] dx dy + Z_2 \iint_{\Delta 2} [ \quad ] dx dy + \dots + Z_n \iint_{\Delta n} [ \quad ] dx dy$$

Now consider a general term of the series

$$Z_P \iint_{\Delta P} \zeta_s(x, y) e^{-j2(\vec{k} \cdot \vec{r}_s)} dx dy$$

If the size of  $\Delta P \gg \ell^2$  where  $\ell$  is the correlation length of  $\zeta_s$ , then the integral

$$\iint_{\Delta P} \zeta_s e^{-j\vec{k} \cdot \vec{r}_{xy}} dx dy$$

can also be expressed as a sum over smaller areas

$$\iint_{\Delta P} \zeta_s e^{-j2\vec{k} \cdot \vec{r}_{xy}} dx dy = \sum_m \iint_{\Delta m} \zeta_s e^{-j2\vec{k} \cdot \vec{r}_{xy}} dx dy = \sum_m X_m$$

Each  $X_m$  is a random variable depending on  $\zeta_s$ . By choosing the subpatch area  $\Delta m$  as a square of dimensions  $\ell \times \ell$ , then each subpatch will be independent of the others. Neighboring subpatches would be somewhat correlated, but here we will assume that the effect is small and neglect it. The mean of  $X_m$

$$\langle X_m \rangle = \iint_{\Delta m} \langle \zeta_s \rangle e^{-j2\vec{k} \cdot \vec{r}_{xy}} dx dy = 0$$

since by definition  $\langle \zeta_s \rangle = 0$ .

The variance of  $X_m$

$$\langle X_m^2 \rangle = \iiint \langle \zeta_s(x, y) e^{-j2\vec{k} \cdot \vec{r}} [\zeta_s(u, v) e^{-j2\vec{k} \cdot \vec{w}}]^* \rangle dx dy du dv$$

$$= \iiint\limits_{\mathcal{V}} R_{\zeta}(x-u, y-v) e^{-j2[k_x(x-u) + k_y(y-v)]} dx dy du dv$$

by stationarity of  $\zeta_s$ , i.e., if  $\zeta_s$  is spatially homogeneous.

This fourfold integral depends only on the differences  $x-u$ ,  $y-v$  and will be the same for every subpatch. If the number of subpatches is large enough, then by the central limit theorem,  $\sum_m \chi_m$  is a zero mean Gaussian random variable  $G$ , the same for all patches. Then

$$\bar{E}^s = G \sum_p Z_p = G \cdot S \quad \text{where } S = \sum_p Z_p.$$

This is precisely the form of an SIRV. As the radar beam moves in time, either by scanning or as it is attached to an aircraft, the returns will be samples of a process, an SIRP.

#### CONCLUSION

A radar clutter model derived from electromagnetic principles has been shown to have the proper form of a spherically invariant random process. Characterization of radar clutter as an SIRP from theory may provide an important link to predicting the specific clutter statistics based on terrain topology and material parameters  $(\mu, \epsilon)$  that make up the surface. It should be noted, however, that a more general formulation of a radar clutter model should include volume scattering (by vegetation) as well as surface scattering, and also provide for near normal incidence and near grazing incidence.

## REFERENCES

- [1] Guinard, N.W., Ransone, J.T., Laing, M.B., and Hearton, L.E.: 'NRL terrain clutter study, phase I', Naval Research Laboratory, NRL report 6487, May 1967
- [2] Fay, F.A., Clarke, J., and Peter, R.S.: 'Weibull distribution applies sea clutter'. IEE Conf. Publ. 155 (Radar 77), 1977 pp.101-104
- [3] Trunk, G.V., and George, S.F.: 'Detection in non-Gaussian sea clutter'. IEEE Trans., 1970, **AES-8**, pp. 620-628
- [4] Jakeman, E., and Pusey, P.N. : 'A model for non-Rayleigh sea echo', 1976, **AP-24**, pp. 806-814
- [5] Jakeman, E.: 'On the statistics of K-distributed noise'. J. Phys. A, 1980, 13, pp. 31-48
- [6] Conte, E., Longo, M.: 'Characterisation of radar clutter as spherically invariant random process'. IEE Proc., Vol. 134, Pt. F, No. 2, April 1987 pp. 191-197
- [7] Yao, K.: 'A representation theorem and its applications to spherically-invariant random precesses', IEEE Trans., 1973, **IT-19**, pp. 600-608
- [8] Goldman, J.: 'Detection in the presence of spherically symmetric random processes', 1976, **IT-22**, pp. 52-58
- [9] Brehm, H.: 'Description of spherically invariant random processes by means of G-function', Springer Lecture Notes, 1982, 969, pp. 39-73
- [10] Beckmann, P., Spizzichino, A.: 'The scattering of electromagnetic waves from rough surfaces', New York, Pergamon Press, 1963
- [11] Brekhovskikh, L.M.: 'The diffraction of waves by a rough surface, part I, Zh. Eksper. i. Teor. Fiz. 23, 1952 pp. 275-289
- [12] Brown, G.S.: 'Backscattering from a Gaussian-distributed perfectly conducting rough surface', IEEE Trans., **AP-26**, 3, 1978 pp. 472-482
- [13] Ulaby, F.T., Moore, R.K., Fung, A.K.: 'Microwave remote sensing - - active and passive': Vol II, Addison-Wesley, 1982
- [14] Rice, S.O.: 'Reflection of electromagnetic waves from slightly rough surfaces', Comm. Pure Appl. Math. 4, pp 351-378, 1951
- [15] Burrows, M.L., 'A reformulated boundary perturbation theory an electro-magnetism and its application to a sphere', Can. J. Phys. , vol. 45, pp. 1729-1743, May 1967
- [16] Burrows, M.L., 'On the composite model for rough surface scattering', IEEE Trans. **AP-21**, pp. 241-243, Mar 1973

- [17] Beckmann, P.: 'Scattering by composite rough surfaces', Proc. IEEE, Vol 53, pp. 1012-1015, Aug 1965
- [18] Fung, A.K., Chan, Hsiao-Lien,: 'Backscattering of waves by composite rough surfaces', IEEE Trans. AP-17, 5, Sept 1969
- [19] Bass, F.G., Fuks, I.M., Kalmykov, A.I., Ostrovsky, I.E., Rosenberg, A.D., 'Very high frequency radiowave scattering by a disturbed sea surface, part II', IEEE Trans. AP-16, 5, pp. 560-568, Sept 1968
- [20] Bass, F.G., Fuks, I.M.: 'Wave scattering from statistically rough surfaces', Inst. Radiophy. Electronics, Kharkov, USSR, translated: Vesecky, C.B., Vesecky, J.F. pp. 418-442 Pergamon Press 1979
- [21] Valenzuela, G.R.: 'Scattering of electromagnetic waves from a tilted slightly rough surface', Radio Sci., Vol. 3, no. 11, pp. 1057-1066, Nov 1968



**PHOTOREFRACTION IN  $\text{Bi}_{12}\text{SiO}_{20}$  AND SEMI-INSULATING  $\text{InP:Fe}$**

**Wallace B. Leigh**  
Assistant Professor, Division of Electrical Engineering  
Alfred University, Alfred, New York 14802

Final Report for Work Performed at  
Rome Laboratories, Solid State Directorate  
RL/ERX, Hanscom AFB MA 01731

Under the Faculty Summer Research Program  
May 25 - August 4, 1991

**ABSTRACT**

This is the final report concerning results of a summer research program at Rome Laboratories. Four separate areas of research were investigated. The first areas focused on analyzing the mechanisms of photorefraction in Czochralski (Cz) and Hydrothermal  $\text{Bi}_{12}\text{SiO}_{20}$  (BSO) grown at Rome. Deep defects in the bandgap of BSO were characterized and defect phenomena related to photorefraction measurements in BSO. Electrical conductivity was measured below and above room temperature for different samples of this material. Above RT conductivity indicates a rapid increase in current in BSO that is thermally activated, and appears dependent on deep defects. A deep defect of  $1.3 \pm 0.1$  eV was found and measured in both Cz and Hydrothermal material. Below RT thermally stimulated current (TSC) studies, however, indicate that the concentration of traps of activation energy  $< 0.7$  eV in the Hydrothermal samples is approximately a factor of  $10^3$  smaller than in Cz BSO. At least five different defects were identified in the TSC measurements.

Activation energies for two of these defects were determined from initial rise measurements of TSC data. While a large concentration of shallow traps was always found in Cz material, the identity of the most dominant shallow trap was not always the same. Importance of this data on photoconducting and photorefracting qualities of BSO are given. Suggestions for further studies are also given. Photorefraction itself was investigated as a method of characterizing BSO and semi-insulating InP:Fe. A report of photorefraction in InP:Fe is also given.

## I. INTRODUCTION

Photorefraction is a name given to a non-linear optical effect observed in certain insulators and semi-insulators. Photorefraction is a method of writing volume holograms which are dynamic, i.e. volume gratings which erase after cessation of exposure with time constants which vary from milliseconds to several months. Since it is a non-linear effect, it holds promise for many of the suitable applications of non-linear mixing including image processors; phase conjugation using degenerate four-wave mixing; optical devices such as optical oscillators; and as a technique useful for non-destructive, non contacting materials analysis [1-4].

Photorefraction has been observed in several insulators, mostly ferroelectrics with large electro-optic coefficients and ~~small~~<sup>large</sup> dielectric constants. In this study photorefraction and general defect properties were investigated in cubic  $\text{Bi}_{12}\text{SiO}_{20}$  (BSO) and semi-insulating iron-doped InP (InP:Fe). Introductory studies of BSO single crystals grown at Rome Laboratories at Hanscom AFB are part of a long-term investigation relating defect properties of BSO to its photorefractive properties. The emphasis of photorefraction in InP:Fe

was twofold: as an attempt to demonstrate photorefraction in this material using low power (HeNe) and incoherent (LED) sources, and to later demonstrate the technique as a method of nondestructive and contactless testing of this material.

An optimum photorefractive crystal must meet the following requirements:

- 1) be fairly transparent at the wavelength of interest,
- 2) have a relatively large amount of deep defect centers in its bandgap,
- 3) possess a large electro-optic coefficient/dielectric constant ratio, and
- 4) be a relatively good photoconductor.

A common theory given for photorefraction is the single-carrier model summarized mathematically by Kukharev [5]. Two coherent pump beams enter the sample with an angle of  $2\theta$  between them, as shown in Fig. (1). The single carrier model of what happens as they interact is shown in Fig. (2). Where the coherent pump beams mix they form an intensity interference pattern of alternating light and dark areas that can be modeled as:

$$I(x) = I_0 + I_1 e^{i(\omega t - kx)} + c.c. \quad (1)$$

In the high intensity regions of the interference pattern, the light is absorbed in the crystal as it alters the population of deep traps in the bandgap. In the single carrier model, the light is absorbed by exciting electrons (holes) to conduction band (or valence band). The free carriers diffuse (or drift if an external electric field is applied), to dark areas of the interference pattern where they recombine. The result is a periodic space-charge region, as shown in Fig. 2b. This

space charge region sets up a periodic electric field, the amplitude of which (with no external field) is determined as:

$$E_{sc} = -i \frac{I_1}{I_0} \frac{E_0}{1 + E_0/E_N} \quad (2)$$

$$\text{where } E_N = \frac{e N_A}{\epsilon K} \quad E_0 = \frac{k_B T K}{e}$$

The resulting electric field causes a periodic change in the refractive index from the electro-optic coefficient,  $r$ , defined from

$$\Delta n \propto r E_{sc} \quad (3)$$

The resulting refractive index modulation is modeled

$$n = n_0 + n_1 e^{i(\omega t - Kx)} \quad (4)$$

The result of the mixing of the two pump beams is a refractive index grating having grating spacing given by Bragg's Law

$$\lambda = 2 \Delta \sin \theta \quad (5)$$

The result is a refractive index grating of spacing  $\Delta$  from which a third beam of light (referred to here as the "probe") can diffract in a similar way to which a beam may diffract from a fixed etched grating. The probe beam (shown dashed line in Fig. 1) in this study was always of a wavelength larger than the pump beam, and thus diffracts at a Bragg angle larger from the pump beam as demanded by equation (5). This is the so-called non-degenerate four-wave mixing configuration. The probe beam produces a diffracted fourth beam which can be used as a means of monitoring the grating.

The fraction of the probe beam diffracted depends only on the relative magnitude of the pump beams, and not on their absolute

magnitude. This relative magnitude of the pump beams is what determines the modulation index  $m = I_1/I_0$ .

The time dependence of the index grating is determined by [6]:

- 1) Dielectric relaxation time,
- 2) Diffusion length of the photocarriers,
- 3) Debye screening length.

Which of these, or combination of these parameters is important is dependent upon the intensity of the light and the grating period. When making a grating, the rise time of the diffraction efficiency increases if total writing intensity increases. The writing and erasing rates increase if the angle between the pump beams increases [7]. Erasing times can also be increased with an erase beam, which offers a very short relaxation time by altering the photoconductivity of the sample.

The significance of Eq (2) being imaginary is a  $1/4$  period phase shift between the electric field and the light intensity. This phase shift can be observed by comparing Fig. 2a to 2d. This  $1/4$  period shift allows one pump beam to self-diffract from the grating into the other beam. Thus a power transfer between the two pump beams exists, termed beam coupling. Beam coupling can be observed without the use of the probe beam. For some ferroelectrics, the magnitude of energy transfer in two-beam coupling can approach 100%, one transmitted pump beam emerges extinguished compared to the other.

If a probe beam of the same frequency and polarization as the pump is used incident with one pump beam; in so called degenerate four-wave mixing, then the fourth beam is diffracted at an angle along the direction as the other pump beam. Do to the non-linear nature of the mixing in the crystal this fourth beam is the phase conjugate of the

pump beam it follows along. Due to the possibility of beam coupling explained above, i.e. one pump beam is diffracted into the other, a reflection of one of the pump beams back into the crystal can be used as the third beam as "self pumped" phase conjugation, that is, phase conjugation using only one beam from one laser [8].

In spite of the simplicity of Kukharev's model, it appears that photorefraction in BSO and InP:Fe involves a more complex mechanism. In fact two-carrier mechanisms have been suggested for both BSO, and InP. For BSO, it is apparent that a hopping conduction typical of insulators is involved [9]. Hopping takes place when carriers move between localized centers in the bandgap.

Photorefraction is unique in that creating the non-linear effect does not require high intensity sources. This is because the grating modulation index  $m$  does not depend on the absolute intensity of the pump-beams, but rather on the ratio of pump beam intensities. Thus it is conceivable that in the right materials, it could be observed using available low-power coherent sources. The speed of the effect does, however depend on light intensity.

In this study several samples of BSO grown by Czochralski (Cz) and Hydrothermal techniques were studied in an attempt to compare their material defect properties to the photorefractive efficiency and time constants of the materials. Four Cz samples, and one hydrothermal sample were measured. These samples were:

BSO 26	6-9s purity Cz grown
BSO 45	5-9s purity Cz grown
BSO 55	Cz grown from a low-purity Hydrothermal charge
BSO 72	5-9s purity Cz grown in $\text{Bi}_2\text{O}_3$ rich conditions
BS 40	Hydrothermal grown
BSO 51	5-9s purity Cz irradiated with 50Mrad gamma radiation
Itek Commercial BSO sample	

All samples except BS 40 were yellowish in color. Sample BSO 72 had several inclusions and was thus unsuitable for photorefractive studies. All samples except BSO 26 and commercial sample were cut and polished at Rome for photorefractive measurements. Separate samples were used for high temperature conductivity and low temperature thermally stimulated current (TSC) measurements. Photorefractive measurements were made using non-degenerate four wave mixing to measure diffraction efficiencies and grating decay times. Preliminary results on photorefractive in InP using infrared pumps and probes is also reported. Each section will be reported separately.

## II. HIGH TEMPERATURE ELECTRICAL CONDUCTIVITY

Electrical conductivity measurements of BSO above room temperature was attempted so that trap or traps having thermal activation energies greater than 0.7 eV could be thermally emptied and thus characterized.

### II.A. HIGH TEMPERATURE CONDUCTIVITY EXPERIMENTAL PROCEDURE.

Measurements were made in the dark on samples from  $80^{\circ}$  to  $450^{\circ}\text{C}$ . A field of  $10^2$  to  $10^3$  V/cm was applied constantly for the entire heat and cool cycle, and D.C. current was monitored using a picoammeter. A five hour linear heat rate between  $85^{\circ}\text{C}$  and  $450^{\circ}\text{C}$  was followed by a three hour soak at  $450^{\circ}\text{C}$  and then a five or eight hour cool time. Samples were contacted at two points using either indium or graphoil contacts. Prior to measurements, indium contacts were pressed and then heated on a hot plate until the indium wetted the sample surface.

### II.B. HIGH TEMPERATURE CONDUCTIVITY RESULTS.

Obtaining measurements of the high temperature conductivity of BSO was complicated somewhat by the rapid onset of space charge effects. Space charge effects were not observed below  $400^{\circ}\text{C}$  at the low electric

fields used, but between 300 and 400°C, measured DC currents for BSO samples would increase dramatically, then saturate when 450°C was reached. A typical scan is shown schematically in Fig. 3. Upon cooling, the current would exhibit a hysteresis, and not return to its initial value for several hours. At this point, the sample would be electrically stressed, or polarized. Upon removal of the electric field at 85°, a reverse polarization current of several tens of picoamps would flow for an indeterminable amount of time.

This reverse current could be observed for all samples tested, including hydrothermal BSO. For some samples, including the hydrothermal grown sample, this reverse current could be temporarily quenched with illumination of an unfiltered low-power tungsten light source. Current would resume after cessation of the illumination. Due to the speed of this photoconductive quenching, it was concluded that the polarization was electronic and not ionic in nature.

One of the more noticeable properties of the data in Fig. 3 is its reproducibility in an unstressed sample-and reproducibility from sample to sample. Due to the dramatic (four orders of magnitude) rise in current through the same temperature range, it is likely that there are deep traps emptying contributing carriers which in turn build a space charge effect. If an unstressed sample is measured, the initial rise data from Fig. 3 can be plotted in an Arrhenius plot and the activation energy determined. Figure four shows the Arrhenius plots of a Cz and a Hydrothermal sample. Activation energies of 1.3 and 1.1 eV, were determined, respectively, from a linear regression analysis of the data. Oberschmid [11] also observed an activation energy of  $1.1 \pm 0.07$  eV on unstressed BSO samples. Determining activation energies from this



technique at these temperatures is hampered by the possibility of charge retrapping. This retrapping is also more dominant if the trap of interest is only partially filled, thus some variation in activation energies is to be expected.

It is interesting to note that similar deep defects were observed from the high temperature measurements in both hydrothermal and Cz BSO, suggesting there does exist deep defects in the hydrothermal material which can, at least, be thermally emptied. This does not explain why such traps are not observed by optical absorption. It is possible that such traps may not fill unless slightly heated above room temperature. This hypothesis may be supported by the fact that clear hydrothermal samples were observed to darken when heated on a hot plate under ambient conditions.

Electrical stressing has been observed by Efendiev, et al [10] in BTO and by Oberschmid [11]. Oberschmid determined that stressing was caused by regions of a negative space charge which slowly builds up as electrons are injected from the cathode and eventually the space charge region shunts the electrodes. Efendiev observed that in BTO the effect could last for several months in the dark. No attempt to determine time constants of electrical stressing was attempted in this study.

Sample BSO 45, after . cycles between 85 and 450 at  $10^3$  V/cm, was visibly observed to have bleached due to the treatment. Sample BSO 26 was not observed to discolor, however was a lighter color initially than the other Cz samples. It was concluded that the treatment either removed an impurity in BSO 45 not in BSO 26, or else indium from the contact was drifted into the sample. As Cz BSO intentionally aluminum or gallium doped at Rome is observed to bleach, it may be expected that

It would have the same effect.

## II.C. SUGGESTIONS FOR FUTURE WORK ON HIGH TEMPERATURE ELECTRICAL CONDUCTIVITY

It is felt that the high temperature current measurements could be taken with fields below 10 V/cm, to further minimize the above mentioned space-charge effects. A much lower field and lower heat rate would help in determining the thermal activation energies of any traps participating in the effect.

It was observed that hydrothermal samples could darken when heated above approximately 150°C. Deep defects could account for this change and above temperature optical measurements of this material could be appropriate.

## III. THERMALLY STIMULATED CURRENT STUDIES OF BSO.

In order to characterize the more shallow defects in BSO TSC measurements were carried out. Linear heat rates could be obtained from 77K up to approx 200 K.

### III.A. TSC PROCEDURE.

Samples approximately 1cm X 1cm X 3mm were contacted with rolled indium 1mm X 1cm strips along the edges of one face. These contacts were heated 15 min in air on a hot plate until the indium wetted the surface. The samples were mounted in a cryostat using rubber cement. The samples were cooled in the dark with no field. Once liquid nitrogen temperature had been reached, a field of 100 V/cm was applied and the sample was illuminated with unfiltered light from a 200 watt Hg arc lamp in order to create free carriers to fill traps. Various illumination times from 2 min to 15 minutes were used with no great observable effect on the resulting TSC spectra. Ten minute illumination times were more

typical. After illumination, the samples were then warmed in the dark at a heat rate of 1.5 C/hr.

### III.B. TSC RESULTS

Typical results of TSC scans are shown in Fig. 5. Three groupings of peaks are normally observed for each sample. Fig. 5a shows at least two peaks exist in group labeled "A". Peak(s) A is a major shallow peak of significant concentration which was observed for most Cz samples.

A rough estimation of trap concentration can be determined by assuming all traps were filled by the filling illumination and then measuring the integrated charge under each peak. Results for several BSO samples show in in Table I.

Table I. Measured integrated charge for the peaks found in TSC measurements of BSO samples.  
All concentrations are in  $\text{cm}^{-3}$ .

<u>Sample</u>	<u>Peak A</u>	<u>Peak B</u>	<u>Peak C</u>	<u>Peak D</u>
BSO 45	$1.1 \times 10^{14}$	$10^{12}$	$10^{12}$	$10^{12}$
BSO 72	$7 \times 10^{13}$	$10^{12}$	$10^{12}$	$10^{12}$
BSO 51 Irrad	$4.4 \times 10^{13}$	$10^{12}$	$10^{12}$	$10^{12}$
BSO 55	N/A*	$3 \times 10^{15}$	$5 \times 10^{14}$	N/A
Itek	$6.7 \times 10^{14}$	$2.25 \times 10^{14}$	$4 \times 10^{14}$	$2 \times 10^{14}$
Hydrothermal	$9 \times 10^{10}$	$10^{11}$	$10^{11}$	N/A

---

\*N/A means the peaks were not observed for these samples.

In Table I, it is apparent that a dominant shallow ( $<0.3\text{eV}$ ) trap of concentration greater than  $5 \times 10^{13} \text{ cm}^{-3}$  was observed in all Cz grown BSO. However, comparing Sample BSO 45 and BSO 55, which was grown from an impure hydrothermal charge, it is evident that the dominant shallow trap is different for these two samples. TSC spectra for BSO 45 and 55 are shown in Fig. 5. It was observed that both of these samples were yellowish and were also both photorefracting.

In Table I and Fig. 5b, it is observed that a significantly lower

concentration of traps was found in the hydrothermal sample, although peaks A through C were observed. This sample was not photorefracting.

In section II it was observed that a deep center was observed in all BSO samples, both Hydrothermal and Cz grown. It is interesting that samples which are photorefracting also have large concentration of shallow levels, however it does not seem important exactly which shallow levels are present, as long as at least  $5 \times 10^{13} \text{ cm}^{-3}$  are indeed present. There may be a connection between shallow defect concentration and the optical absorption of BSO. If there are not enough shallow traps in the Hydrothermal material, deep defects present may be above the Fermi level and are thus empty at room temperature and are not emptied by room temperature optical absorption. These traps could therefore not participate in photorefraction.

From the initial current rise of Peak A in sample BSO 45 and Peak B in sample BSO 55, thermal activation energies were determined as described by Milnes [12]. Initial slope of the data was fitted to

$$I = I_0 \exp\left(-\frac{\Delta E}{k_B T}\right)$$

For these measurements, the TSC scans were repeated at higher gains compared to those used to record figure 5 in order to obtain accurate readings of the initial rise in thermal currents. Arrhenius plots for these two peaks are shown in Fig. 6. Activation energies determined from a least-squares analysis gave an activation energy of 0.17 eV for Peak A and 0.215 eV for Peak B.

It is also interesting to note in Table I that the Itek sample has higher integrated charge under Peaks B through D as compared to those grown in Rome.

It was also observed that BSO 55 was not a good photoconductor

under the un-filtered Hg light, at any temperature. However sample BSO 55 is still a reasonable photorefractor, as observed in Section IV. Thus a study relating photoconductivity to efficiency and speed of photorefraction is probably warranted. The effects of photoconductivity quenching via the tungsten light as observed above may have some effect on photorefraction. It has been observed that infrared illumination does increase photorefraction efficiencies [14].

### III.C. SUGGESTIONS FOR FURTHER TSC WORK.

The ease of the TSC measurements have indicated that this could be a good technique for monitoring BSO samples grown at Rome. It is felt that a definite correlation could be made between the purity of the starting Cz charge and the type and height of the TSC peaks observed. It may prove that in order to make hydrothermal samples photorefracting, shallow impurities may have to be added to change the way charge is compensated by deep states. A lack of shallow impurities may move the Fermi level in such a way as to empty deep traps that are normally filled with electrons in Cz material.

A cryostat suitable for TSC needs to be developed. Such a cryostat, where heat rates are more linear, can be warmed through room temperature, and where heat rates can be varied, is needed to determine more accurate trap activation energies.

### IV. PHOTOREFRACTION OF BSO.

The samples were next measured for photorefraction efficiency, as measured by non-degenerate four wave mixing, and grating time constant. These measurements were undertaken to determine the feasibility of using photorefraction to characterize BSO grown at Rome Labs.

#### IV.A. PHOTOREFRACTION PROCEDURE.

A schematic of the photorefractive system is shown in Fig. 5. For BSO measurements, the pump beam was a 466 nm line from an Argon laser and a 633 nm HeNe beam was used for the probe. The pump beams to the sample from the Argon laser were created from a cube beam splitter. The sample was placed in the crossed pump beams which made an angle of 14° between one another. The probe beam from the HeNe was bounced from mirror "A" onto the sample. A target was used to mark the relative positions of pump beams and probe beam. The target was first placed in the pump beams, at a distance of 12 inches in front of the sample, and the positions where the pump beams intersected the target was marked. The target position (12 inches) is also the position where mirror "A" is placed. The Bragg angle for the pump beams can be calculated from measuring the distance between pump beams marked on the target. The Bragg angle for the probe can next be calculated and its relative position on the target can be marked. This is the point where the probe beam will incident the mirror. The HeNe laser is mounted such that the incident spot on mirror "A" could be x-y translated. The HeNe beam is translated until the red light meets the mirror at the point marked on the target. This would bring the probe within a few degrees of its Bragg angle, and with minor adjustments of mirror "A" from that point the diffracted beam could be maximized. A 1.35  $\mu\text{m}$  probe beam from a 44 microwatt LED was attempted but was difficult to observe photodiffraction due to the large Bragg angle. At a large angle, the errors on the target are large, and the probe beam is easily moved off the sample with even minor corrections of mirror "A" to the incident angle. Both pump and probe lasers were vertically polarized.

For efficiency and time constant measurements, the diffracted probe was monitored with a silicon photodetector. The incident probe beam was modulated at 30 Hz and a lock-in detection system was used. Neutral density filter of O.D. - 2 or 4 was inserted in the incident probe beam to insure that the probe would not erase the grating and that the silicon detector would not be exposed to the high irradiance of the diffracted beam.

#### IV.B. PHOTOREFRACTION RESULTS.

Efficiency measurements were roughly the same for all samples, at  $2.5 \times 10^4$  for 5mW probe attenuated by O.D. - 2 filter. When the sample was rotated  $90^\circ$ , efficiencies were 25% higher in one orientation than the other. However, these measurements are not satisfactory for the reasons stated below.

Photorefractive measurements were hampered by several problems which made reproducibility of measurements difficult. Reliable measurements could only be taken when the system was in perfect alignment. Since different efficiencies are observed for the two different grating vectors that are possible in cubic crystals, the samples must be rotated and measured for both directions. Since only two samples were cut on all six sides, a rotator had to be constructed which would rotate a sample. Unfortunately, the rotator also moved the sample in and out of where the beams crossed, and subsequently once rotated the system had to be realigned. Time restraints did not allow for the construction of a better rotator. A rotator needs to be designed which would allow rotation of the sample about a fixed point without changing the angle of incidence to the sample surface.

The second more serious problem for efficiency measurements is the

effect of the probe beam on erasing the grating. Probe beam of 633 nm, when attenuated by increasing O.D. filters, gave increasing diffraction efficiencies. With the probe attenuated by O.D. = 4, the efficiencies would go up a factor 10. With O.D. greater than 4 the diffracted beam could not be observed with the detection system used.

Time constants for three samples shown in Fig. 8. These decays were measured in the dark. Incident probe intensity was attenuated by O.D. = 2 filter for these measurements. However, these measurements suffer the same problems as efficiency measurements, meaning, the effect of the probe beam in erasing the grating cannot be accounted for. If the probe is attenuated to very low levels, then the ambient light in the room would erase the grating, and the ambient light is not reproducible for day to day measurements.

#### IV. SUGGESTIONS FOR FURTHER WORK.

It is felt that the effects of the probe intensity could be negated in the time constant measurements in the above experiments if the grating was purposely erased with light of known wavelength and intensity. However, the time constants in doing so are shorter than the response time of the system used in the above measurements. In the experiments above, the time constant of the measurement system was determined by the output filter of the lock-in, or the highest speed of the strip chart recorder used, typically 1m sec.

The erasure-under-light time constant will depend on the photoconductive dielectric response time of the sample when exposed to that light, which is a material parameter and thus may give some information concerning material quality.

Efficiency measurements could be made with the stronger 488 nm



pump and an infrared probe which is not absorbed as strongly in BSO as the red. If a very sensitive detection system were used, efficiency could be measured with increasing attenuation of the probe until the measured efficiencies saturated, and thus the effects of the probe could be discounted.

#### V. PHOTOREFRACTION OF InP.

Photorefraction experiments were attempted on InP:Fe grown at Rome labs. Samples 1 X 1 X 0.8 cm were cut and two faces along the short direction were polished. These faces were [110], the other directions being [110] and [100]. An electric field of 3kV was applied across the [100] direction. For this purpose ohmic contacts of thin indium were rolled onto the sample sides and annealed at 400 C for 1.5 minutes.

The pump beam in the InP experiments was a 1.15  $\mu\text{m}$  2 mW HeNe laser. The probe used was a 44  $\mu\text{W}$  1.35  $\mu\text{m}$  LED. The LED was collimated using a X10 objective lens. The laser was expanded using a X10 objective, and 2 inch F.L. convex lens and spatially filtered. The same set up for photorefraction of BSO was used. The same cube beam splitter and mirrors were used for InP as for BSO photorefraction.

#### V.B. RESULTS AND SUGGESTIONS FOR FUTURE WORK.

Unfortunately, photorefraction was never observed in the InP samples. Difficulties in aligning infrared beams were encountered. A serious problem with the InP is not knowing the correct material parameters for photorefraction. This is a basic "chicken and egg" problem. Optical density of several InP samples varied from 0.6 to 1.0 for roughly the same sample thickness, and the relation between absorption and defect quantity in InP is not that clear cut.

It is suggested that a sample of InP that is a known

photorefractor be obtained. With this sample, the low-power system used in this study could be tested. If photorefraction were successful at this power level with this sample, samples grown at Rome could then be tested. Otherwise, many samples grown at Rome would have to be tested, initially with higher power lasers than those used in this study.

## REFERENCES

1. P. Günter, J.P. Huignard, eds., Photorefractive Materials and Applications, volumes 1 and 2, Springer-Verlag, Heidelberg (1988).
2. J. Feinberg, "Photorefractive Nonlinear Optics," Physics Today, October, 1988, p. 46.
3. N.V. Kukhtarev, J. Feinberg, D.M. Pepper, "The Photorefractive Effect," Scientific American, 1989.
4. G.C. Valley, P. Yeh, Guest eds., J. Opt. Soc. Amer. B. Special Issue on Photorefractive Materials, Effects and Devices 5, 1682-1821 (1988).
5. N.V. Kukhtarev, V.B. Markov, S.G. Odulov, M.V. Soskin, V.L. Vinteskii, "Holographic Storage in Electrooptic Crystals. I. Steady State" Ferroelectrics, 22, 949 (1979).
6. G. Pauliat, J.M. Cohen-Jonathan, M. Allain, J.C. Launay, G. Rosen, "Determination of the Photorefractive Parameters of  $\text{Bi}_{12}\text{GeO}_{20}$  Crystals Using Transient Grating Analysis," Opt. Comm 59, p. 266 (1986).
7. J. Feinberg, D. Heinman, A.R. Tanguay, Jr., and R.W. Hellworth, J. Appl. Phys. 51, 1297 (1980).
8. A. Yariv, Optical Electronics, Fourth Edition, Holt, Rinehart, and Winston (1991) Chapter 18.
9. A.E. Attard, "Theory and Origins of the Photorefractive and Photoconductive Effects in  $\text{Bi}_{12}\text{SiO}_{20}$ ," J. Appl. Phys. 69, p. 44 (1991).
10. M. Efendiev, V.E. Bagiev, A.Kh. Zeinaly, M. Grandolfo, and P. Vecchia, "Deep Localized Centers in Sillenite-Type non Linear Crystals," Ferroelectrics, 43, 217 (1982).
11. R. Oberschmid, "Conductivity Instabilities and Polarization Effects of  $\text{Bi}_{12}(\text{Ge},\text{Si})\text{O}_{20}$  Single-Crystal Samples," Phys. Stat. Sol. (a) 89, 657 (1985).
12. A.G. Milnes, Deep Impurities in Semiconductors, John Wiley 1973 p. 236.
13. A.A. Kamshilin, M.G. Miteva, "Effect of Infra-red Irradiation on Holographic recording in Bismuth Silicon Oxide," Optics Communication 36 429 (1981).

## LIST OF FIGURES

1. Schematic of the non-degenerate four-wave mixing experiment involved in observing the photorefractive effect.
2. Model illustrating the photorefractive effect using the single-carrier model.
3. Schematic of temperature & current for above room-temperature electrical conductivity measurements, showing hysteresis in current.
4. Arrhenius plot of the initial rise of high temperature conductivity for different BSO samples. Cross-sectional areas were the same for all samples at  $0.25 \text{ cm}^2$ .
5. TSC measurements of a) Cz BSO sample BSO 45, and b) Hydrothermal sample BS 40, c) Cz BSO sample BSO 55, grown from Hydrothermal charge.
6. Arrhenius plots of the initial rise of TSC peaks "A" of sample BSO 45 and "B" BSO 55.
7. Schematic of Photorefraction experiment.
8. Dark photorefraction time constants for samples commercial BSO and BSO 26. Each decay is repeated with the sample rotated  $90^\circ$ .

PUMP  
BEAM

PROBE BEAM

2θ

PUMP  
BEAM

SAMPLE

Fig 1

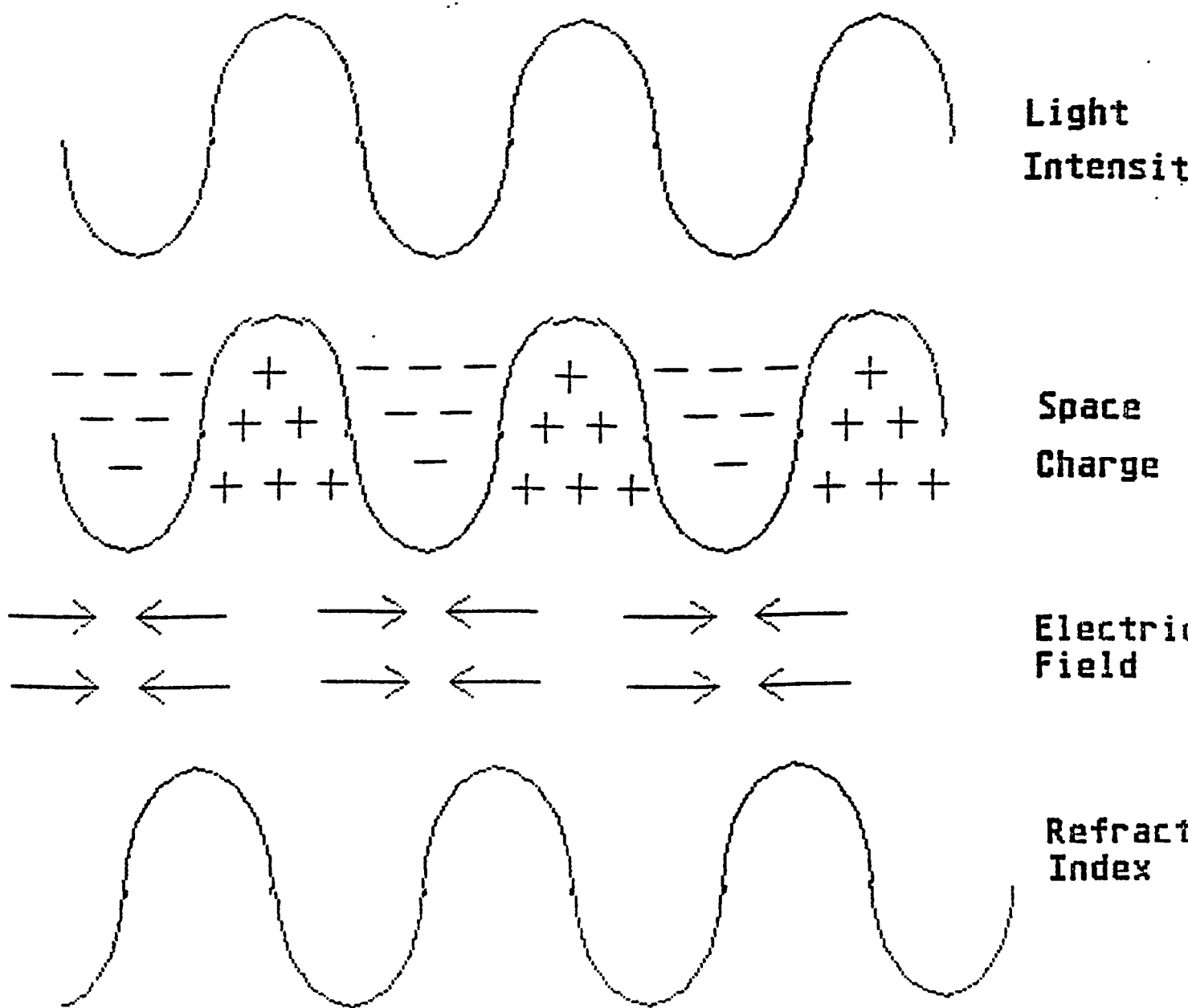


Fig 2

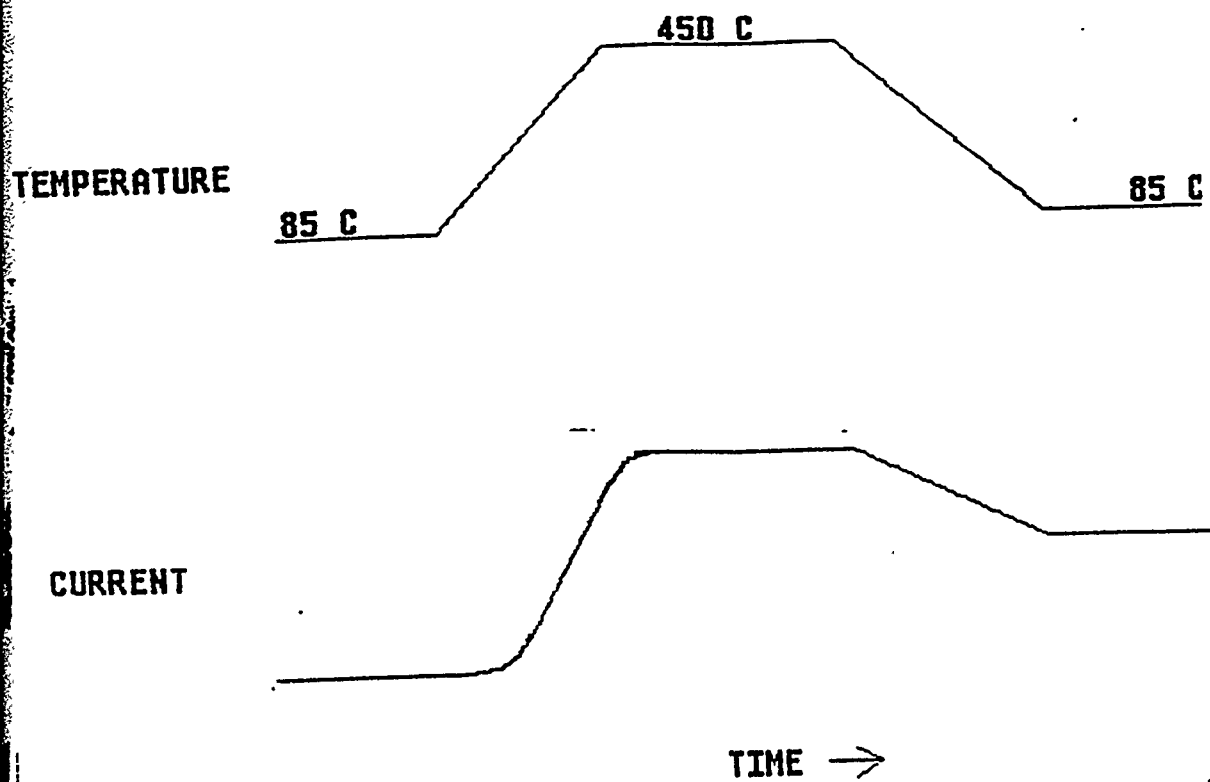
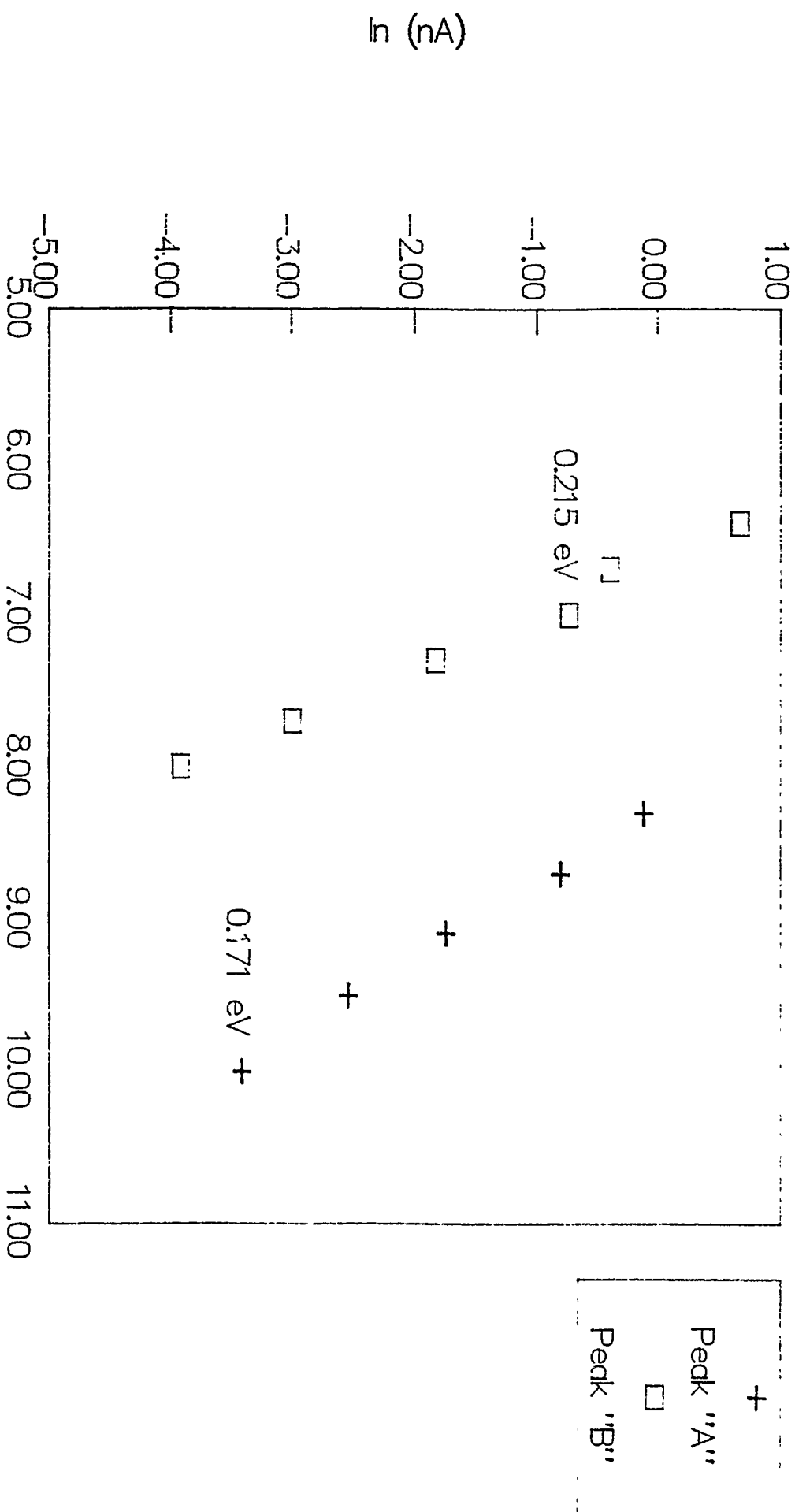
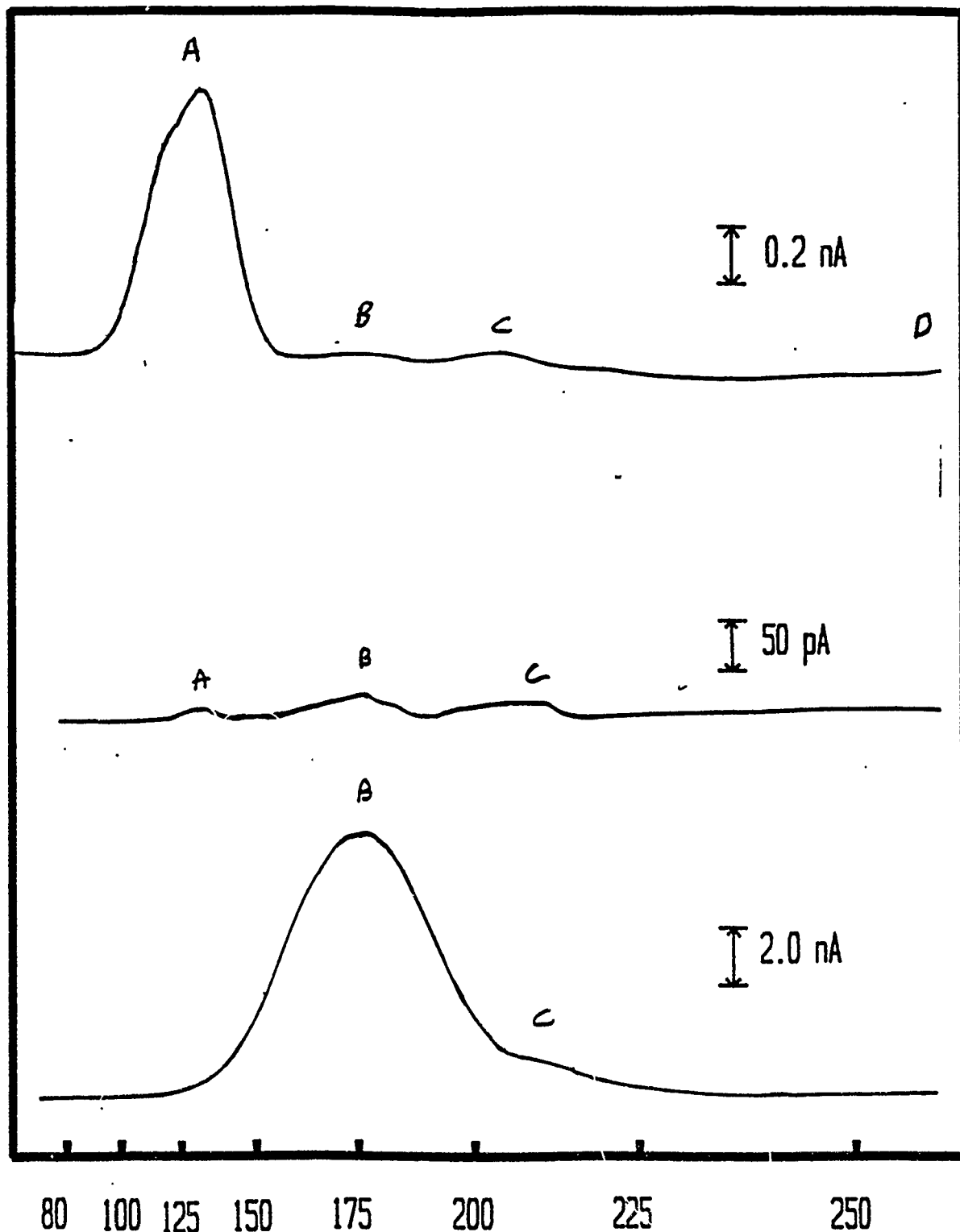


Fig 3

## Arrhenius Plots of TSC Data



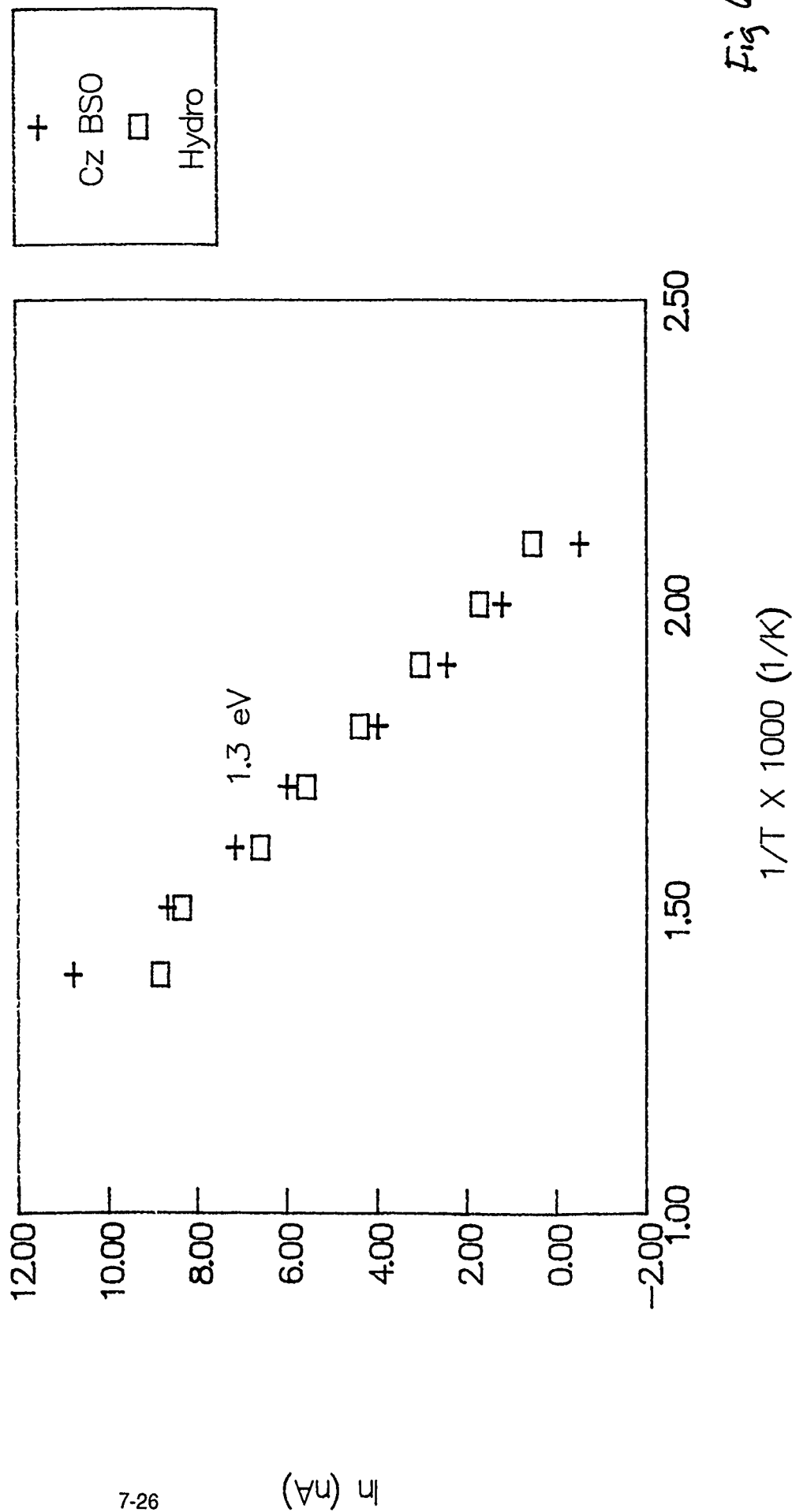




Temperature (K)

# Arrhenius Plots of High Temp Peaks

## Hydrothermal and Czoehralski BSO



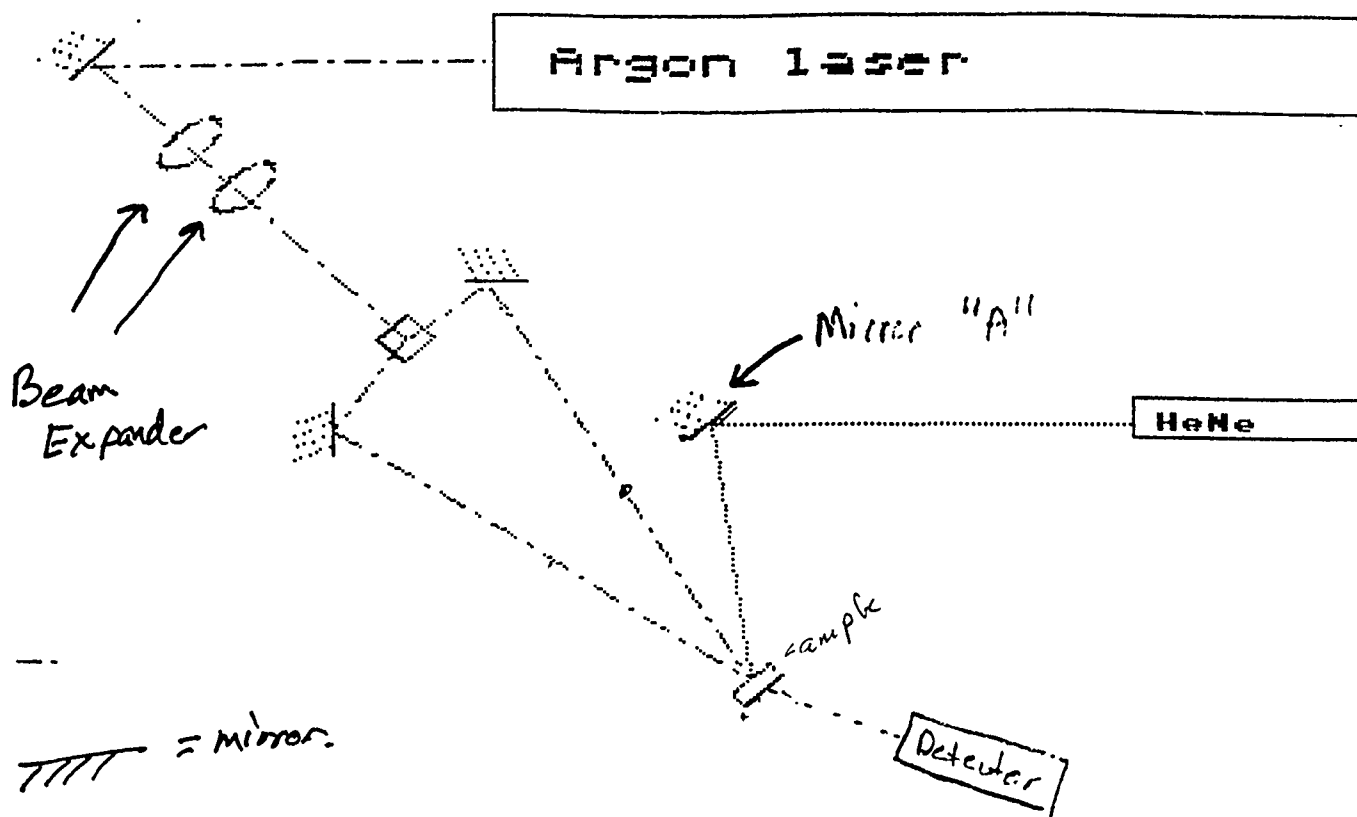
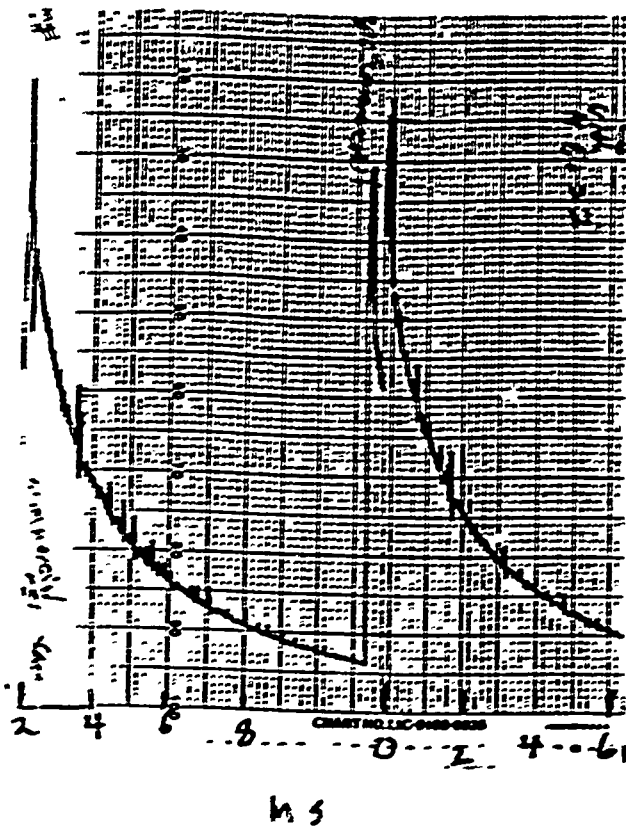
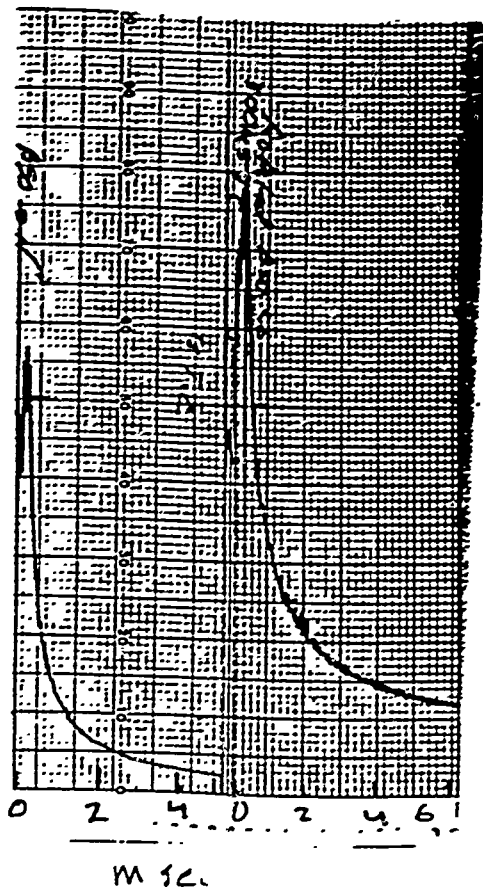


Fig 7

(a)



(b)



# **DEVELOPMENT OF A METHODOLOGY FOR EXTRACTING SEMANTIC RELATIONS FROM DEFINITIONS**

**Elizabeth D. Liddy  
School of Information Studies  
Syracuse University  
Syracuse, New York 13244-4100  
liddy@mailbox.syr.edu**

## **ABSTRACT**

A methodology was developed for the task of automatically extracting semantic information from dictionary definitions by means of Relation Revealing Formulae (RRF) which are based on lexical, syntactic and semantic regularities in Longman's Dictionary of Contemporary English and which make use of the additional information provided on the machine readable tape. Those verb definitions which use 'to make' comprise the sample on which the generic Constant Comparative research paradigm was applied for the development of the specific steps of a new methodology for extracting semantic relations. RRF which successfully disambiguate the senses of 'make' as used in definitions were developed and a plan for further testing of the RRF and refinement and application of the methodology is proposed.

## **1. Introduction**

The goal of my summer research conducted at Rome Labs was to further develop and refine a methodology for extracting semantic knowledge from a machine-readable dictionary (MRD). The methodology should have generalizability to the extraction of semantic knowledge from other large textual databases. The semantic representation of the dictionary definitions which this methodology would produce could be used in two very different types of applications.

Firstly, the methodology provides a reasonable approach to the task of developing reasonably-sized lexicons for NLP systems. The creation of semantic representations of texts is a major hurdle to be overcome in the development of text-based intelligent systems (TBIS), whether these systems be traditional information retrieval systems or whether they be expert systems making use of naturally occurring texts as one of their

knowledge-bases. The lack of reasonably sized, semantically rich lexicons is posing one of the major bottlenecks in TBIS development (Lenat et al, 1986). This dearth of lexicons is due to the absence of a method for the efficient acquisition of semantic knowledge of the world. The automatic or semi-automatic acquisition of the lexicon is considered a critical factor in determining how widespread the use of natural language processors will be in the next few years (Velardi et al, 1991). The knowledge for use in such systems must be acquired and represented in such a manner that its implicit semantics are then available for the necessary inferencing of the host system. Knowledge acquisition, carried out by knowledge engineers, is the most commonly used approach to acquiring this knowledge, but it has proven time-consuming and difficult to contain. The methodology which I have been developing this summer provides one plausible alternative to this inefficient process. It is called *knowledge extraction* (Kwasnik, Liddy & Myaeng, 1989), and is a means whereby pre-existing textual sources are processed in as automatic a manner as possible with the goal of extracting the knowledge encoded in natural language repositories.

Secondly, the semantic knowledge which this methodology would extract from the MRD could be used to produce a humanly "explorable vocabulary" consisting of nodes and relationally-labelled arcs for use as a browsing tool for individuals who are attempting to learn about a new domain of knowledge. Usually, such initial learning is disadvantaged by an individual's relative unfamiliarity with the semantic structure of the concepts in a new field of interest. One could envision a graphical semantic representation (e.g. semantic network) being a useful tool; presenting the concepts of the field and specifying the semantic relations among them. This explorable vocabulary could serve to clarify an individual's original notions; to provide appropriate terminology for further investigation of the topic; and to suggest richer connections between concepts which had

not yet occurred to the novice.

Machine Readable Dictionaries (MRDs) form an excellent and readily available textual source for performing semantic knowledge extraction because MRDs: 1) are culturally validated sources of much of our commonplace knowledge of the world as accepted by native language speakers over many years; 2) provide raw data which is unbiased by a particular individual's beliefs, views, knowledge or experience, and; 3) span a wide range of subject areas, thereby providing a shallower, but more complete coverage of the language than would be available in other corpora.

For my research, I am using the 1987 version of Longman's Dictionary of Contemporary English (LDOCE). The 1977 version of this MRD has been the object of a substantive body of work (Alshaw, 1987; Boguraev & Briscoe, 1987; Wilks et al, 1987) which has revealed the potential richness of the data encoded on it. LDOCE is particularly appropriate for my work because it was developed to be used as a learner's dictionary by non-native English speakers and, to accommodate this use, is based on a defining vocabulary of approximately 2000 words. Most of the definitions are written using only these terms and any word used in a definition that is not on this list is indicated by a type-font code. This controlled defining vocabulary contributes to the tractability of my current research.

## 2. Background

The current research is based on earlier investigations I have conducted into the nature and content of the definitions in machine readable dictionaries, theories of semantic relations, sublanguage analysis and some joint planning of a humanly explorable vocabulary (Kwasnik, Liddy & Myaeng, 1989) and requires some basic explanation of those areas.

## 2.1 Semantic Knowledge in MRDs

Meaning is commonly defined as consisting of *concepts* and the *relations* among these concepts. In the English language, concepts are most frequently expressed in the form of noun phrases while relations among concepts are commonly expressed by either verb phrases or prepositions. There are, of course, exceptions to these regularities, but the linguistic practice is consistent enough for its assumption to permit development of a regularized approach to the automatic extraction of meaning from English text.

The *concepts* present in dictionary definitions are easier to recognize automatically than the semantic *relations* that exist among concepts, because the language used in dictionary definitions only implicitly reveals these relations. For example, a definition may not expressly state "the purpose of X is Y", but, according to my hypothesis, there is a delineable set of phrases such as "X is an instrument for Y-ing" in which the phrase "is an instrument for" reveals the PURPOSE relation between X and Y. I define *relations* as properties that hold between two or more entities. The entities may be people, events, objects, situations, characteristics, actions, places, values, etc. Relations define the nature of the interaction, dependency, influence or simply co-occurrence which holds between the entities.

Although semantic relations have been of research interest in a variety of disciplines such as psychology, cognitive science, ethnography, and linguistics (Evens, 1988), no consensus has been reached on a definitive set of semantic relations. I have been proceeding inductively to uncover all the semantic relations present in dictionary definitions, based on the belief that the most appropriate set of relations cannot be determined a priori, but is discoverable within the corpus being analyzed (White, 1988). Empirical tests of applications using the representations created from my analyses will reveal both the coverage and the utility of the set of relations developed.



## 2.2 Semantic Relations in Definitions

Figure 1 provides some examples of how semantic relations are expressed in LDOCE definitions. The terms following the parentheses exist in relation to the term being defined, in the relation stated in capital letters within the parentheses. The name of the relation within the parentheses was added by me. The phrases in each definition presented in bold print are the lexical clues used in the definition which suggest the nature of that particular relationship. For example, in the definition of *diet*, the GOAL relation exists between *diet* and 'become thinner' and is indicated by the phrase **In order to**.

---

*depress* - to **make** [CAUSAL] less active or strong  
*smooth* - **having** [CHARACTERISTIC] an even surface  
*pitcher* - a player **who** [AGENT] throws the ball toward the batter  
*stab* - to **strike** forcefully **with** [INSTRUMENT] the point of something sharp  
*diet* - to eat according to a special diet, esp. **In order to** [GOAL] become thinner  
*shoot* - an area of land **where** [LOCATION] animals are shot for sport  
*nail file* - a small **instrument** with a rough surface **for** [PURPOSE] shaping finger nails

Fig. 1: Sample Definitions from LDOCE

---

## 2.3 Sublanguage Analysis

The linguistic basis of this approach to automatic extraction of knowledge from MRD derives from sublanguage theory (Sager et al, 1987) which has shown that within a particular text-type, predictable syntactic and semantic regularities develop over time, and these regularities are used consistently. Sublanguages develop amongst a group of language users who share a common purpose and who must repeatedly produce instances of a particular type of text (Kittredge & Lehrberger, 1982; Grishman & Kittredge, 1986). Since lexicographers are a specialized group working on a common task, the

linguistic features of dictionary entries can be analyzed using sublanguage methodologies which have shown themselves successful in other text-types consisting of brief, fairly standardized utterances (Liddy et al, 1989; Liddy et al,1991).

Dictionaries use relatively conscribed ways to convey a definition. This fact has been noted and exploited by other lexical researchers who refer to the "words or phrases that are used frequently in definitions" (Ahlsweide & Evens, 1988) as defining formulae (Smith, 1981; Ahlsweide, 1985). My work builds upon these earlier efforts, but distinguishes itself from other lexicographic research on MRDs in that I am developing more complex Relation-Revealing Formulae, rather than the simpler defining formulae.

#### 2.4. Relation-Revealing Formulae (RRF)

RRFs couple a semantic relation to the particular linguistic features of definitions which reliably indicate the presence of that semantic relation. The difference between defining formulae and RRFs is that the former make use of only lexical or, less frequently, syntactic string matching, whereas RRF are capable of capturing and representing a broader scale of complexity in the regularities of definitions. The three levels of RRF which I am developing are presented in Figure 2 and each is followed by a sample RRF template of that level.

---

**LEXICAL RRF:** Require straightforward lexical string matches wherein the exact word or phrase which has been established as a lexical clue to the relation is matched in the definition. More complex LEXICAL RRF require the recognition of non-adjacent phrase patterns.

**RRF: LOCATION = 'where' + <      >**

**SYNTACTIC RRF:** Consist of either sequences of words of specified parts-of-speech or combinations of lexical string matches and specified parts-of-speech and takes advantage of the grammar codes available on the LDOCE tape.

**RRF: RESULTANT = 'to make' + <adjective>**

**SEMANTIC RRF:** Depend on information about semantic features of words (e.g. animacy) in the formulae being available in the dictionary entry box codes. It is typical for a single Semantic RRF to include lexical and syntactic matching as well as semantic matching.

**RRF: EXPERIENCER = 'a'/'an' + <animate noun> + 'who' + <stative verb>**

Fig. 2: Three Levels of Relation-Revealing Formulae

---

### 3. Methodology

Figure 3 provides my long-range research plan for the development and testing of the RRF methodology. The results of this work will later be integrated with a neural net approach for detecting relations which together will form an Intelligent Semantic Relation Assigner (Liddy & Paik, 1991b).

---

1. Maximum Coincidence Searching of LDOCE definitions
2. RRF template development
3. Semantic relation tagging of one-half of LDOCE
4. Semantic Relation Formulae testing on remaining one-half of LDOCE
5. Evaluation of Relation Revealing Formulae results
6. Development of an Intelligent Semantic Relation Assigner (ISRA)

Fig. 3: Research Overview

---

The main effort in the preliminary development of the methodology consists of the labor-intensive, rigorous, intellectual analysis of large sets of definitions for the inductive discernment of semantic relations which are predictably indicated by the various levels of linguistic regularities described above. This is the methodology development which I have been focusing on this summer and consists of the following steps:

1. Computationally analyzing one-half of the definitions in LDOCE using the Maximum Coincidence Search procedure to exhaustively identify all re-occurring phrases, including non-adjacent patterns.
2. Establishing the specific semantic nature of the relations revealed by these linguistic regularities.
3. Coupling of the highest level of abstraction of these linguistic regularities with their appropriate relation in RRF templates.

### 3.1. Maximum Coincidence Search (MCS)

The Maximum Coincidence Search technique, which was developed for the study of LDOCE by Woojin Paik (Liddy & Paik, 1991a) searches for, and extracts, the longest possible string match (whether terms are adjacent or not) within a specified textual unit, and creates sets of these patterns (i.e. coincidences of terms) while tagging the text for the position in which each pattern occurred.

MCS is being used during RRF development to extract all possible coincidences of terms from text. Referring to the example presented in Figure 4, one can see how the MCS technique works. Processing the LDOCE tape using MCS detects the reoccurring pattern of adjacent and non-adjacent combinations of the two words 'act' and 'of': a total of 750 times in a variety of patterns. The first column presents ten observed instances taken directly from LDOCE. The second column lists the generalized surface **lexical** patterns existing in these instances, while the third column presents five **lexical/syntactic** patterns underlying the ten instances. These five patterns can then be abstracted to the final RRF developed for the semantic relation: ACT. The template capable of correctly recognizing all 750 occurrences of the ACT Semantic Relation, as indicated by any combination of 'act' and 'of', is presented at the bottom of Figure 4. All occurrences, when individually analyzed did indicate this semantic relation and none other.

<u>OBSERVED INSTANCES</u>	<u>LEXICAL PATTERN</u>	<u>LEXICAL/SYNTACTIC PATTERN</u>
<i>an act of admonishing</i> <i>the act of accepting</i>	an act of ...ing the act of ...ing	ART + <i>act of</i> + PRES.PART
<i>a single act of breathing</i> <i>an official act of forgiving</i> <i>the sound or act of firing</i>	a __ act of...ing an __ act of ...ing the __ act of ...ing	ART + __ + <i>act of</i> + PRES.PART
<i>an act of secretly escaping</i> <i>the act of purposely staying</i>	an act of __...ing the act of __ ...ing	ART + <i>act of</i> + __ + PRES.PART
<i>an act or state of boiling</i> <i>the act or process of converting</i>	an act or __ of ...ing the act or __ of ...ing	ART+ <i>act</i> + __ + <i>of</i> + PRES.PART
<i>an example or act of not being...</i>	a __ or act of __...ing	ART+ __ + <i>act of</i> + __ + PRES.PART

RRF for ACT = ART + [ ] + *act* + [ ] + *of* + [ ] + PRESENT PARTICIPLE

Fig. 4: Sample MCS Patterns

Although this is a trivial example, it exemplifies precisely how MCS allows me to examine the machine-readable tape of LDOCE efficiently in order to detect all possible variations of word patterns and to uncover from amongst the surface variety the basic phrasal patterns, which then undergo intellectual analysis for the inductive discernment of the semantic relation indicated by each basic pattern and to develop templates which can be used by an algorithm to detect a particular semantic relation no matter how the basic phrasal pattern is varied.

### 3.2. Establishing the RRF Methodology

The establishment of a methodology for developing RRF was begun with an analysis of verb definitions in LDOCE. One reason for this choice was to facilitate the research of Michael McHale of Rome Labs, whose development of a principle based parser (McHale,

1991) requires knowledge of the thematic roles (e.g. agent, experiencer) of noun phrases for the purpose of eliminating ill-formed syntactic structures. The information as to thematic roles is not explicitly available in any of the coded fields in LDOCE but RRF offer one means of determining which thematic roles occur with each verb. I also considered verbs a useful group to analyze, since research to-date on the simpler defining formulae has focused on either noun or adjective definitions, with little attention to verb definitions.

Most definitions provide a genus term for the word being defined and then some differentia that distinguish the word being defined from other classes of the same genus. Since verbs are usually defined with a verb as their 'genus' term, a frequency count was made of the individual verbs from one-half of the verb definitions in LDOCE, and showed that 'make' occurred 529 times; 'cause' occurred 402 times and 'be' occurred 380 times. Given these figures, the choice was made to begin with 'make' and 'cause' for development of RRF. The work has proceeded on both of these verbs, with some productive overlaps in the RRF, but only the development of RRF based on 'make' will be herein reported, as the RRF for 'cause' are not refined to my satisfaction.

By using MCS, I retrieved all 2, 3, 4, and 5 word (contiguous or noncontiguous) word patterns which contain 'make'. Given that the development of RRF begins as an inductive process, this is the first step to be done. Results of the MCS produce data on which the simplest level, lexical RRF are based, but, in addition, given appropriate analysis, the retrieved data will also suggest what the more complex syntactic and semantic RRF might be, and how information available on the machine-readable tape might be used in developing this type of RRF.

As an example, retrieving the 5 word patterns containing 'to make' includes the following patterns and their frequencies:

to make (something)	35	times
to make ( )by	30	"
to make (someone)	27	"

Although none of these 5 word patterns may be indicative of a particular semantic relation, they do suggest some lexical patterns to be looked for in shorter phrases and also begin to suggest some of the semantic variations of 'to make' as used in LDOCE definitions. In that the goal of RRF development is to devise a means whereby the semantic relations contained in definitions can be explicated, it is essential that the possible senses of frequently occurring verbs such as 'make' can be disambiguated. For instance, it would seem intuitive that the sense of 'make' would be different when followed by 'something' rather than by 'someone' and that semantic representations of these definitions should reflect these differences.

By consulting the definition of the word 'make' itself as provided in LDOCE, one sees that 'make' has three basic senses, which can be briefly paraphrased as:

1. to PRODUCE
2. to PERFORM
3. to CAUSE or CAUSE TO BE

This information can be entered as the hypothesis into the inductive-deductive loop

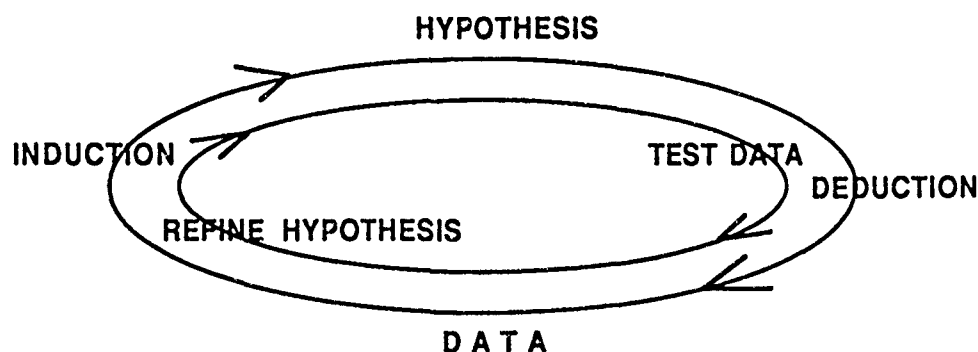


Fig. 5: Constant Comparative Methodology

---

of the Constant Comparative methodology (Glaser & Straus, 1967) as seen in Figure 5. This is the basic research paradigm I am following in this effort.

Next, with these hypothesized senses in mind, I used the second function of the MCS program to look at the sets of definitions which contain a searched pattern. Analyzing these sets of definitions, the syntactic and even the semantic patterns begin to suggest themselves. For example, there are many definitions which are of the basic form:

'to make' + ADJECTIVE

where the great semantic variety of the adjective does not matter in terms of the predictable relation between the word being defined and that adjective. For example, the definitions in Figure 6 all reveal the CAUSE TO BE relation with the adjective that follows 'to make', although in some instances it might not follow it directly.

---

constrict - to make [CAUSE TO BE] narrower, smaller, or tighter  
consummate - to make [CAUSE TO BE] (a marriage) complete by having sex  
convey - to make [CAUSE TO BE] (feelings, ideas, thoughts, etc.) known  
defuse - to make [CAUSE TO BE] less dangerous or harmful  
exhilarate - to make [CAUSE TO BE] (someone) cheerful and excited

Fig. 6: Definitions containing CAUSE TO BE sense of 'make'

---

In addition to identifying that 'to make' reveals the CAUSE TO BE relation between the verb being defined and the following adjective, I have also identified the adjective or adjectives as being in the RESULTING ATTRIBUTE relation to the verb being defined.

Figure 7 shows these relations.

---

constrict - to make [CAUSE TO BE] narrower [RES.ATT], smaller [RES.ATT], or tighter [RES.ATT]  
consummate - to make [CAUSE TO BE] (a marriage) complete [RES.ATT] by having sex  
convey - to make [CAUSE TO BE] (feelings, ideas, thoughts, etc.) known [RES.ATT]



deafen - to make [CAUSE TO BE] deaf [RES.ATT], esp. for a short time  
 defuse - to make [CAUSE TO BE] less dangerous [RES.ATT] or harmful[RES.ATT]  
 exhilarate - to make [CAUSE TO BE] (someone) cheerful [RES.ATT] and excited  
 [RES.ATT]

Fig. 7: CAUSE TO BE and RESULTING ATTRIBUTE relations

---

Further analysis allowed me to delineate the semantic nature of the other entities in the CAUSE TO BE schema. These are PATIENT for all human nouns and AFFECTED OBJECT for all other nouns. In addition, the MEANS relation is revealed quite predictably in 'to make' definitions by a phrase beginning with 'by' at the end of the definition. Figure 7 shows the same definitions with these additional relations identified. Whether a noun is human or not can be determined by accessing the Box Codes which are available on the LDOCE tape but not in the printed version of the dictionary and indicate either the semantic nature of the word being defined or the semantic restrictions on the subject and object of the verb being defined or the acceptable noun that can be qualified by the adjective being defined. This is essential knowledge needed for developing Semantic Level RRF.

---

constrict - to make [CAUSE TO BE] narrower [RES.ATT] smaller[RES.ATT], or tighter  
 [RES.ATT]  
 consummate - to make [CAUSE TO BE] (a marriage) [AFF.OBJ] complete [RES.ATT] by  
 having sex [MEANS]  
 convey - to make [CAUSE TO BE] (feelings, ideas, thoughts, etc.) [AFF.OBJ] known  
 [RES.ATT]  
 deafen - to make [CAUSE TO BE] deaf [RES.ATT], esp. for a short time  
 defuse - to make [CAUSE TO BE] less dangerous [RES.ATT] or harmful[RES.ATT]  
 exhilarate - to make [CAUSE TO BE] (someone) [PATIENT] cheerful [RES.ATT] and  
 excited [RES.ATT]

Fig. 8: CAUSE TO BE, RESULTING ATTRIBUTE, AFFECTED OBJECT, PATIENT & MEANS

---

#### 4. Results

Processing of the verb definitions containing 'to make' in one half of LDOCE produced two types of useful results. Firstly, the RRFs themselves, which could successfully differentiate among the various senses of 'to make' and also reveal the semantic roles of other constituents in the definitions, and secondly, the particular steps of the methodology used were coded sufficiently that the methodology can now be taught to research assistants so that further development of a semantic lexicon based on LDOCE can be produced for use in a range of NLP systems.

##### 4.1 RRF for 'to make'

The RRF for the remaining two senses of 'to make' were developed following similar procedures. This effort was aided by the fact that the secondary RRFs developed as by-products while developing RRF for the first sense of 'to make' were usable in other situations as well. Figure 9 presents the basic senses of 'to make' which were evident in the definitions, along with a sample definition containing each of these senses.

- 
1. to make [ a ] = (to produce a)  
*crack* - to make a sudden loud sharp sound
  - 2.a. to make [ b ] = (to cause to be b)  
*disquiet* - to make anxious
  - b. to make [ c ] [ b ] = (to cause c to be b)  
*annoy* - to make someone a little angry or impatient, esp. by repeated troublesome actions or attacks
  3. to make [ c ] [ d ] = (to cause c to d)  
*abase* - to make (esp. oneself) lose self-respect

Fig. 9 : Four senses of 'to make'

---

These four senses of 'to make' can be detected by syntactic level RRF because each of the four patterns can be matched on part of speech information, as seen in Figure 10. These

syntactic level RRF require the RRF applicator to consult the online LDOCE file to determine the parts of speech of the words in the definitions.

---

1. to make [ a ] = (to produce a)  
RRF (PRODUCE) = 'to make' + noun phrase
- 2.a. to make [ b ] = (to cause to be b)  
RRF (CAUSE TO BE) = 'to make' + adjective
- b. to make [ c ] [ b ] = (to cause c to be b)  
RRF (CAUSE TO BE) = 'to make' + noun phrase + adjective
3. to make [ c ] [ d ] = (to cause c to d)  
RRF = (CAUSE) = 'to make' + noun phrase + verb

Fig. 10: RRF for detecting senses of 'to make'

---

In addition, semantic information that is available in the Box Codes on the machine-readable LDOCE can also be included in the RRF if a fuller conceptual representation of the definitions is desired. For example, in Figure 11, if the noun phrase that fills the A

---

1. RRF [PRODUCE] = 'to make' + noun phrase (A)
- 2.a. RRF [CAUSE TO BE] = 'to make' + noun phrase (C) + adjective (B)  
    - or -
- b. RRF [CAUSE TO BE] = 'to make' + adjective (B)
3. RRF [CAUSE] = 'to make' + noun phrase (C) + verb

where:

A = RESULT CONCRETE if Box Code 1 = S, J, N  
or  
A = RESULT ABSTRACT if Box Code 1 = T  
B = RESULTING ATTRIBUTE if Box Code 1 = T  
C = PATIENT if Box Code 1 = Q, P, A, H, B, D, M, F  
or  
C = AFFECTED OBJECT if Box Code 1 = S, J, N, I, L, G

Fig. 11: Fully Realized RRFs for Verbs Defined by "to make"

---

slot in RRF 1, has either Box Code S, J, or N on the LDOCE tape, then that noun phrase is the CONCRETE RESULT of the verb being defined. Otherwise, if the noun phrase has Box

Code T, that noun phrase is the ABSTRACT RESULT of the verb being defined. This is exactly the level of refined semantic representation that is needed by complex NLP systems for which this methodology will be used to develop lexicons.

#### 4.2 RRF Methodology Development

The goal of this summer research effort was to develop a methodology for extracting semantic relations from dictionary definitions. In outline form, the methodology which I have described in procedural detail above, is summarized in Figure 12.

- 
1. Decide part of speech to begin with.
  2. Run frequency distribution of vocabulary used in definitions of this part of speech.
  3. Choose individual words to begin analysis with.
  4. Run Maximum Coincidence Search on these words.
  5. Use Constant Comparative Methodology to iterate between data and hypothesized RRF.
  6. Search for lexical patterns.
  7. Retrieve definitions containing these patterns.
  8. Determine semantic relation being conveyed.
  9. Study fuller sets of definitions to determine whether RRF can be generalized to a syntactic or semantic level RRF.
  10. Test RRF on other half of LDOCE.
  11. Revise RRF and retest.

Fig.12: Steps in RRF Methodology Development

---

#### 5. Conclusions

The successful production of both syntactic and semantic level RRF which can disambiguate the individual senses of 'to make' in dictionary definitions and produce semantically tagged representations of verbs which are defined by 'to make' suggests that this methodology will provide the necessary tool for developing automatic procedures for extracting semantic relations from text.

## 6. Future Work

The current effort has resulted in the successful development of RRF for 'to make' and the specification of a methodology for developing RRF which will produce representations of text in which the implicit semantics are made explicit. There are three possible routes to be followed in the continuation of this work:

1. Empirical testing of 'to make' RRF on the second half of LDOCE.
2. Generalization of 'to make' RRFs to similar verbs.
3. Training Research Assistants to continue RRF development under my supervision.

It is hoped that all of these avenues can be pursued immediately through funding by the Research Initiation Program. It is essential that the research be pursued, for as Velardi et al (1991, p. 157) have stated: "automatic or semi-automatic acquisition of the lexicon is a critical factor in determining how widespread the use of natural language processors will be in the next few years".

## 6. References

- Ahlsweide, T. (1985). A linguistic string grammar of adjective definitions from Webster's Seventh Collegiate Dictionary. In: Humans and machines: Proceedings of the 4th Delaware Symposium on Language Studies. Norwood, NJ: Ablex.
- Ahlsweide, T. & Evens, M. (1988). Generating a relational lexicon from a machine readable dictionary. International Journal of Lexicography, 1 (3), 214-237.
- Alshawi, H. (1987). Processing dictionary definitions with phrasal pattern hierarchies. Computational Linguistics, 13 (3-4), 195-202.
- Boguraev, B. & Briscoe, T. (1987). Large lexicons for natural language processing: Utilising the grammar code system of LDOCE. Computational Linguistics, 13 (3-4), 203-218.
- Evens, M. (1988). Relational models of the lexicon: Representing knowledge in semantic networks. Cambridge University Press.
- Glaser, B & Strauss, A. (1967). The discovery of grounded theory: Strategies for qualitative research. NY: Aldine Pub.

- Grishman, R. & Kittredge, R. (1986). Analyzing language in restricted domains. Hillsdale, NJ: Lawrence Earlbaum Associates.
- Kittredge, R. & Lehrberger, J. (1982). Sublanguage: Studies of language in restricted semantic domains. Berlin: de Gruyter.
- Kwasnik, B. H., Liddy, E. D. & Myaeng, S. H. (1989). Automatic knowledge extraction from dictionary text: Project development. Syracuse, NY: CASE Center Technical Report #8911.
- Lenat, D., Prakash, M. & Shepherd, M. (1986). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. AI Magazine, 65-85.
- Liddy, E. D., Jorgensen, C. L., Sibert, E. & Yu, E. S. (1989). Processing natural language for an expert system using a sublanguage approach. Proceedings of the 52nd Annual Meeting of the American Society for Information Science.
- Liddy, E.D., Jorgensen, C.L., Sibert, E. & Yu, E.S. (1991). Sublanguage grammar in natural language processing. Proceedings of RIAO '91 Conference. Barcelona.
- Liddy, E. D. & Paik, W. (1991a). Automatic recognition of semantic relations in text. Informatics 11. London: ASLIB.
- Liddy, E. D. & Paik, W. (1991b). Automatic semantic relation assigner: Preliminary work. In Natural Language Learning Workshop: IJCAI-91. Sydney, Australia.
- McHale, M. (1991). Principle based parsing using a machine readable dictionary: Abstract of dissertation topic. (unpublished manuscript).
- Sager, N., Friedman, C., & Lyman, M. (1987). Medical language processing. Reading, MA: Addison Wesley.
- Smith, R. (1981). On defining adjectives: Part III. Dictionaries: Journal of the Dictionary Society of America. 3, 28-38.
- Velardi, P., Pazienza, M. T., & Fasolo, M. (1991). How to encode semantic knowledge: A method for meaning representation and computer-aided acquisition. Computational Linguistics. 17(2), pp. 153- 170.
- White, J. S. (1988). Determination of lexical-semantic relations for multi-lingual terminology structures. In Evens, M. (Ed.). Relational models of the lexicon: Representing knowledge in semantic networks. Cambridge: Cambridge University Press.
- Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T. & Slator, B. (1987). A tractable machine dictionary as a resource for computational semantics. In: Proceedings of the Workshop on Natural Language Technology Planning. Blue Mountain Lake, NY: 1-27.

# **APPLICATION-BASED UTILITY EVALUATION: A DISCUSSION LEADING TO ASSESSING THE ROLE OF END-USERS IN THE EXPLOITATION OF VIRTUAL REALITY TECHNOLOGY**

**Michael S. Nilan, Ph.D.**

## **INTRODUCTION**

Systems, whether computerized or not, with or without artificial intelligence, incorporating advanced display technology or not, etc. can be seen as nothing more than a set of procedures and technical capabilities linked together to solve a particular problem or set of related problems. This is as true of a bureaucratic standard operating procedure as it is of a knowledge based distributed group decision support system. Since the advent of computers (and especially in the last ten years), the kinds of problem solutions that we have been implementing in an automated fashion are becoming increasingly sophisticated as are the kinds of display technologies (e.g., multimedia, virtual reality) and peripherals, network architectures, etc. that the system designers employ to deliver solutions in a systematic manner.

As application and information systems become more advanced, they are also becoming more complicated. Some of the concerns of this advancement and its associated complexity are obvious, like increases in training overhead required for users to learn and understand the system. These concerns then become desired goals for system design but solutions are by no means obvious in the current state-of-the-art of system design. Some implications are not as obvious, like evaluating a system for its use of appropriate technology in delivering solutions. This is no longer as straightforward as it once was.<sup>1</sup> Assessment of isolated technology (i.e., not incorporated into a specific system) tends to be in terms of its capability, e.g., processing speed, resolution, transmission speed, etc. The evaluation metrics that are currently employed for system evaluation are oriented towards user performance (and are therefore indirect measures at best anyway) and have been borrowed from education and psychology without significant change. If one assumes that these indirect measures (e.g., retention, accuracy, time on task, demographics, personality

variables, etc.) have been appropriate for simple learning theory and physiological perception, it does not necessarily follow that they are also appropriate for the increased demand on end users or the increased sophistication of the problems so addressed. In particular, these inherited metrics cannot directly assess the complex cognitive behaviors associated with the specific problem solving domain as implemented in any particular system (which includes the particular technical enhancements employed). The increased sophistication and complexity of current and future systems calls for a new look at evaluation.

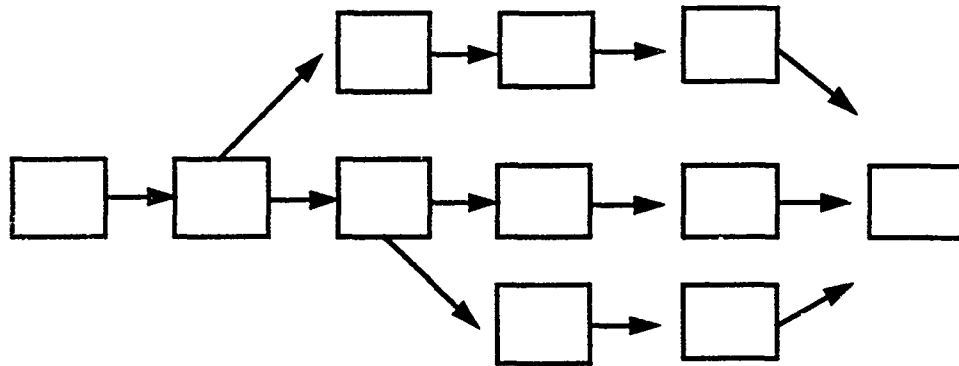
The purpose of this paper is to describe an example of an assessment approach that directly addresses the evaluation issue of the utility of system features (both hardware and software) that are employed (or might be employed) in terms of the specific problem the system is designed to solve. Such application-based utility evaluation procedures have been called for (e.g., Dervin & Nilan 1986) and have begun to be developed in information system contexts (e.g., Nilan 1992; Nilan & Hert 1991), but this paper represents the first attempt to set forth this logic in a more generalized manner, especially in a manner that also allows for concurrent assessment of appropriate technologies (i.e., display devices, pointing devices, tactical feedback mechanisms, etc.) in the virtual reality (VR) context. It is important to note at the outset that the purpose here is to augment current assessment/evaluation procedures rather than to displace them. In other words, the approach presented here is seen as an addition that, when used in conjunction with existing techniques, will facilitate decision making for appropriate technology use in system design and system evaluation.

## **SYSTEMS, SYSTEM DESIGN, AND SYSTEM EVALUATION**

Figure 1 shows a generic description of the notion of a system.

Central to this notion of a system is the particular problem to be solved. This is in keeping with so-called "information resource management" (IRM) strategies currently being implemented in many organizations. Take a manufacturing process as an example of a problem for which we want to design a system. The first step is to define the process in terms of its constituent activities in a time order that reflects the current practice. So, we begin with specifying a product which is to be the goal of the manufacturing process. Let's call it a widget. After we have designed this widget, we must then purchase the raw



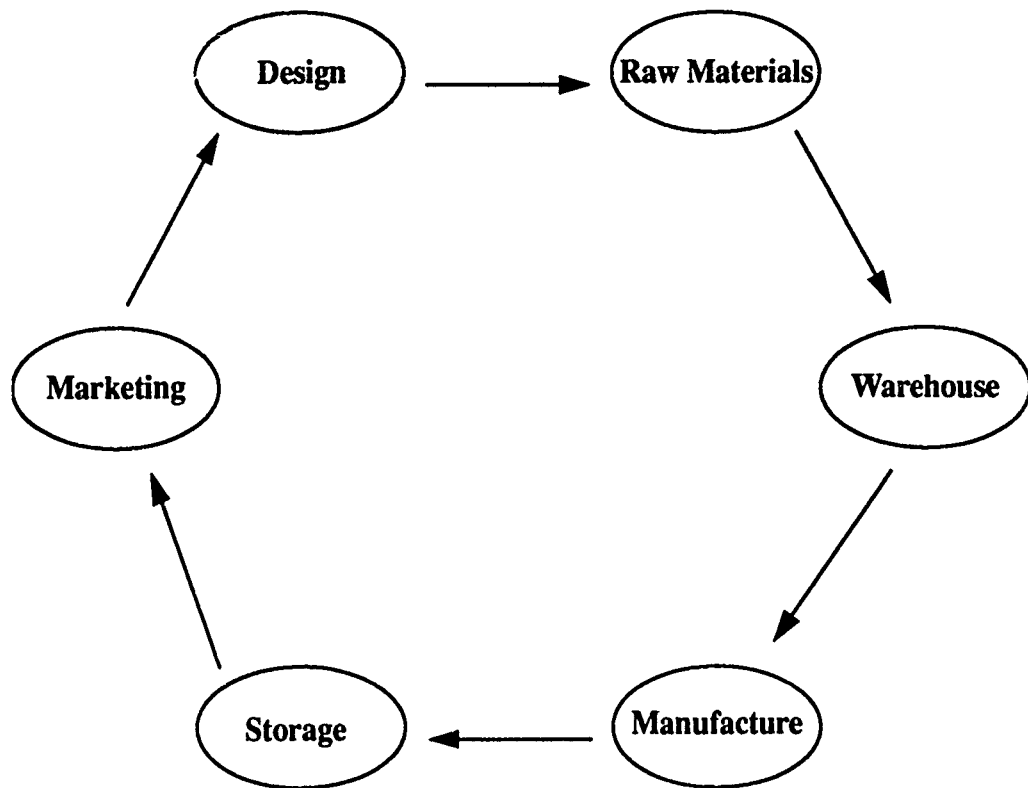


**Figure 1: System** = A set of procedures designed to solve a particular problem. Applies to bureaucratic procedures (e.g., standard operating procedures, forms), the organization of information (e.g., Library of Congress classification system), as well as computerized applications including display technology, interfaces, functional capabilities of the application software and any information associated

materials necessary for producing it, warehouse them until they are needed and transport them to the manufacturing location. At the manufacturing location, we have a sub-system that actually produces the widget, which is then transported to another warehouse to await its sale to a customer. The sale of widgets is carried out by a marketing sub-system that handles such things as advertising, shipping and feedback from the widget purchasers, which ultimately effects the design of the widget. Then, the cycle starts all over again.

Figure 2 graphically illustrates this cycle.

Now that we have documented the process of widget manufacturing, it is possible for an information system designer to examine and evaluate the entire cycle, including various technologies in terms of their ability to provide more efficient manufacturing. So the designer might try to design an information system that allows us to employ “just-in-time” technology. This technology allows us to functionally link sales demands for widgets directly to our purchase of raw materials and to the numbers of employees we employ in the manufacturing sub-system process. What this means for our cycle is that we only order enough raw materials for what we are going to manufacture at any particular time which is in turn linked to the demand for widgets. We have been able to cut out of the cycle the warehousing and storage steps as well as having more efficient use of labor during the actual manufacture. The savings to our organization from a (relatively) simple information



**Figure 2: Hypothetical widget manufacturing cycle.**

system will probably be substantial without any loss of quality of the product or its availability to consumers on demand.

This is an example of a simple system that manages a physical entity, i.e., a widget. It is clear in this example that the evaluation of the just-in-time technology is intimately tied to the problem at hand, i.e., manufacturing widgets. The evaluation criteria employed pertain to efficiency (i.e., time and money) associated with the manufacturing cycle technology itself. Let's take a more complicated example, one that deals with a cognitive process to illustrate the proposed application-based utility approach to evaluation of technology, a decision support system that helps consumers decide what kind of new car to purchase.

This example deals with cognitive activities that do not necessarily have any inherent time dependency associated with individual behaviors, nor is there an established "correct" way to approach the problem. As in the widget example, we first try to define the task at hand. Standard knowledge acquisition procedures for knowledge based or expert

systems would call for interaction with a new car expert (see Agarwal 1988) to establish the criteria associated with purchasing a new car (i.e., things to look for in a new car) and the rules for employing these criteria (i.e., if you have children, then you will need a 4-door or a wagon or a van, etc.). Typical methods involve asking the expert to specify what should be looked for in the purchase of a new car. The answers in this example are exactly what we might expect. The expert tells us we should look for low price, safety, low maintenance costs, good resale value, etc. In fact, the resulting range of requirements is probably the same as we would get if we asked a few non-experts on the street. These requirements would then be built into a computer system that presents the new car buyer with a series of screens that present individual questions and "if...then..." decision rules to assist in the decision making. However, there are some problems that emerge with the resulting system when we evaluate it.

The knowledge base might be evaluated by comparing several experts' determinations of the sets of criteria and rules although this is seldom done in practice (see Kannan 1991 as an example of the use of a single expert for both the knowledge base as well as the subsequent evaluation). The experts might agree to a degree such that we are convinced that from the experts' perspective, the system has sufficient coverage. However, when we look at the use of the system by end users, several surprises occur. First, it takes the users quite a bit of time to learn how the system works. Several walk away from the system because they don't want to be bothered (after all, they are only buying a car). This phenomenon is, of course, not captured by our formal evaluation procedures except perhaps where we look at demographic and/or personality measures and note that those who walked away from the system did not have a "sufficient" technical background. We interpret this as meaning that this system is only for technical people or for those who will "pay" the training overhead to use the system. Other users who did take the time to learn how the system worked walked away before they had finished the entire decision making scenario. They told us it took them too long to use it.

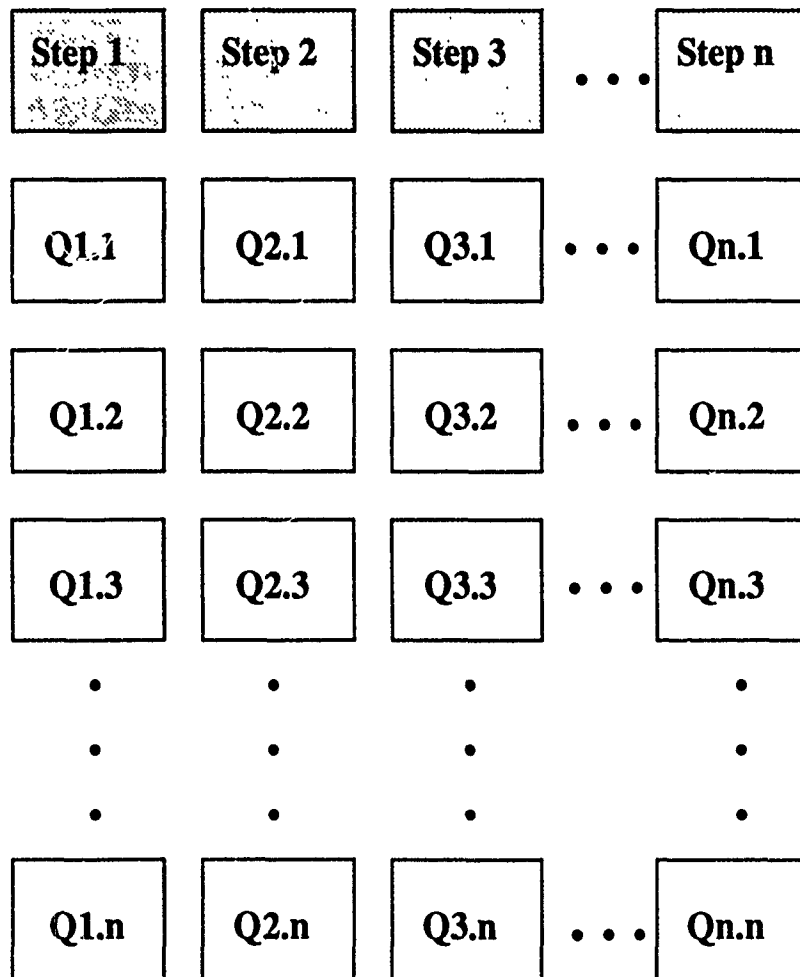
Of those who took the time to learn how to use the system and finished the decision making scenario, we can compare their purchase with what the system told them they *should do* and we find discrepancies. These discrepancies are apparent when we use an "accuracy-" type evaluation criterion where the expert system's determination is seen as the "correct" response and the users' deviations from this

response are their scores on the measure. We may also try to find out from users what criteria actually were behind their decisions, although this is almost never done except in the context of decision support system "satisfaction" measures (see Dervin & Nilan 1986 for a discussion of this category of system evaluation criteria). With these traditional satisfaction measures we find out that users were not satisfied with the advice that the experts gave, discounted the system's advice and therefore, deviated from the "correct" decision. We may even ask the users what they were dissatisfied with (another uncommon tactic) and find that there were several criteria that they considered important that the system did not address that are "intangible." We (the clients for the system) might therefore conclude that the system works well (and probably won't even question its implementation on a computer), but only for those people who have a technical background, who take the time to learn how to use the system, and then follow the system's advice. We do not come away from this evaluation experience with any insight into the technology used, with the possible exception of cost, and we have virtually no data on how to change the system so that it is more useful to users. We might try some intuitive solutions, like putting several questions/rules on a single screen to cut down on the time it takes for a user to run through the scenario. We might even adopt a rapid prototyping technique to involve users in our redesign, but basically, we are stuck with the particular implementation that was originally developed. Although if we go back after a year or so and compare the users' satisfaction with their new car purchases, we will find little difference in the variance between those who followed the system's advice and those who did not. So ultimately, what good is the system?

Let us take the same problem and illustrate an application based utility approach to the design and evaluation of the decision support system. There will be several differences that will be pointed out after the description is presented.

Given that the decision of what to look for in a new car is not a process that involves any physical phenomena and given that there is no inherently "correct" way to go about making this kind of decision, the proposed approach would use a different technique for determining what the nature of the task is. First, recent car buyers would be asked to describe their experiences, from the time they first started to think about a new car until they had actually purchased it, as a series of steps. Then, for each step in the process, the buyers would be asked (for example) what questions they had, whether they got answers to their

questions, what source had the answer, etc. We would end up with a matrix-like description of the activities (or steps) in their car purchasing behavior in conjunction with some of the cognitive information concerns associated with each step. Figure 3 illustrates this activity-information matrix for one such buyer.



**Figure 3:** An abstract activity by information matrix with questions that the user had specified at specific steps as the operationalization of the information component of the matrix. This is essentially a graphic, time sensitive representation of a problem solving experience.

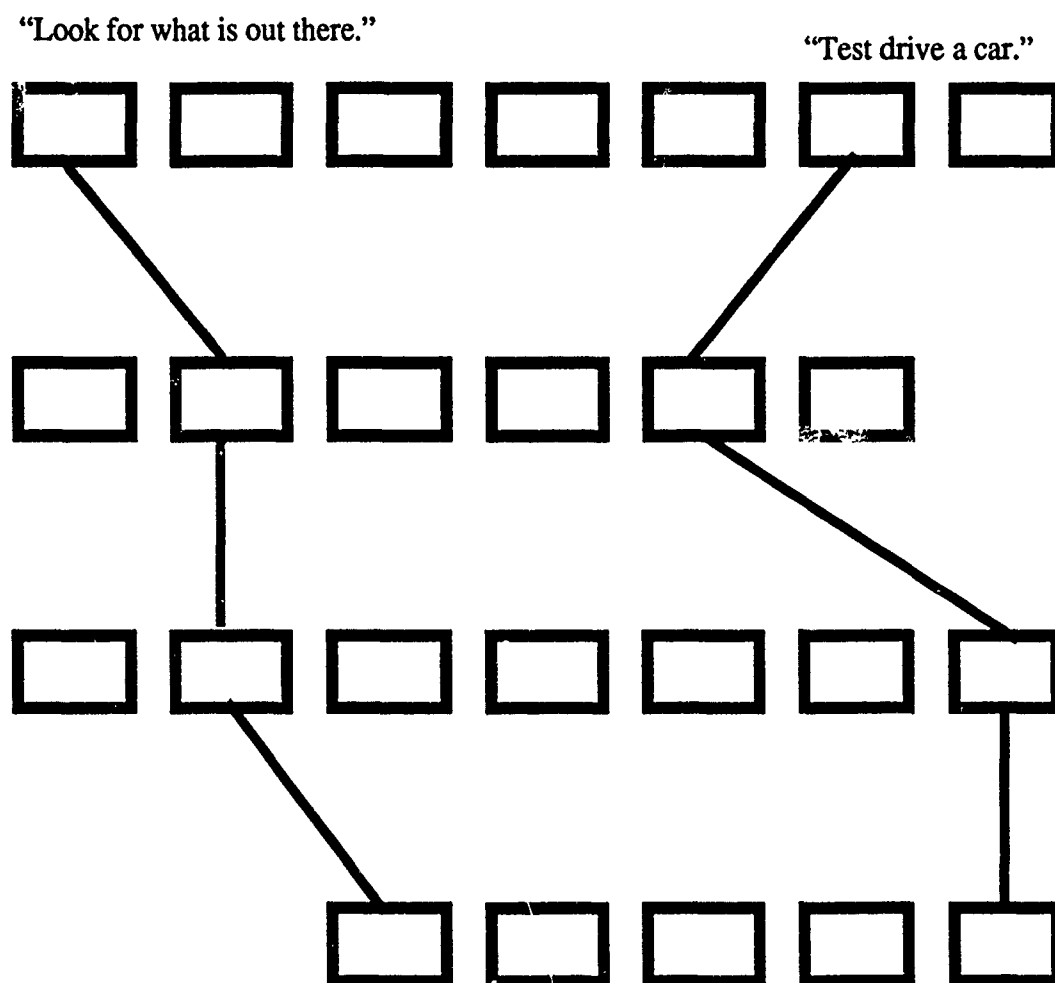
This process would be repeated for a cross-section of recent buyers, ranging from young to old, from expert to novice, rich to poor, etc. (the actual number of people depends

upon the nature of the problem being addressed and the variance in perceptions across the end user population although it is always far fewer than individual differences approaches to system design would suggest<sup>2</sup>). The next step in the process is to design a model of the new car buying problem solving process by aggregating and summarizing across individual car buyer experiences. The model developing process involves simply looking for similarities in the steps in individual accounts in two different ways: first, in the description of the activities (i.e., the actual behavior) and second, in the ordinal time relationship between these similar activities. Figure 4 illustrates this logic and procedure. In the case of car buying, for example, we might find that very early in the process, buyers would go out and look at new cars to see what was out there and begin to classify cars according to price. For system design purposes, there are two striking aspects to this finding: first, virtually all of our new car buyers did this (albeit some of them only in newspapers and magazines) and second, most of them did this early in the problem solving process. We *will* find similar patterns later on in the process (e.g., taking a car for a test drive). The model of the task of buying a new car that we will be able to create by this process will represent a user based view of the task (as opposed to the arbitrary listing of criteria and rules in the expert based view) that is a dynamic representation of the actual problem solving process (analogous to the manufacturing cycle used in the first example). Note also that the information needed for an individual to move through this process is contained in the information side of the matrix and can also be incorporated into the model (see Nilan 1992 for a more detailed description of this analytic process).

The resulting model provides us with a model of the task (i.e., a knowledge base) for which we are designing a system. It incorporates a wide variety of criteria associated with the process that are tied to specific points in the dynamic problem solving process (as opposed to simple lists in the first example). We also note that the range of criteria in this version is much broader than the expert-derived list, for example, the color of the car is important because you don't want to buy a dark car if you live in a hot region since it retains excessive heat in the summer. Now, as in the widget example, we can begin to make decisions about what kind of system we should design.

We note that the process occurs over a long period of time and at several different locations so we need to be sure that users can access our DSS at various times and at various locations. If computer networking was available and most of our users had personal

computers with modems, it would probably be possible to make this system available over a network. Neither of those assumptions is warranted however, but we may be able to make the system available in places where people are thinking about this problem (i.e., in car



**Figure 4:** A hypothetical example of several user descriptions of buying new cars showing the overlap of activities (“look for what is out there” and “test drive a car”) as well as the ordinal, temporal relationship between the two overlaps (“looking” early in the process and “driving” late in the process).

showrooms) if there are computers in those locations. In this hypothetical case a pamphlet is probably a more appropriate technology. Even though it is an inefficient format for updating, it is a format that will allow the user to access it over time and at different

locations. The format is also useful to users because they can keep track of their progress and decisions by writing on the pamphlet. They need no training to use the system because they already know how to read, understand, etc. the set of *a priori* skills necessary for effective use of our system now.

Our model also provides us with a means of organizing the criteria and rules associated with the user based description of the task, i.e., in a temporal sequence which matches the perceptions of how new car buyers approach the problem. Criteria that play an important role early in the process are located "earlier" in the pamphlet so the user accesses the DSS according to where s/he is in the problem solving process (as opposed to the earlier arbitrary listing). (NOTE: If we had decided to use a computerized format for our DSS, these features could also be accommodated such that the user would not have to learn very much to access the system's knowledge base because its organization would be exactly the same).

The evaluation of our system (the "appropriate" technology evaluation was dealt with above), begins with an examination of the coverage of the knowledge base, which is addressed in the knowledge acquisition and analysis phases. By sampling across variables that would seem to account for variance in users' perceptions of the problem solving process (e.g., new buyers versus experienced buyers), we begin to find that talking to more people does not give us any new insight into the task definition process. Once we have begun to experience redundancy in activities and information across pertinent sampling variables, we can stop. While it is probably true that our range is not 100% complete, it was evaluated in much the same way that the expert-derived range (i.e., redundancy among experts versus redundancy among end users). Further, our range is broader than the expert-derived range and includes criteria that users see as important (e.g., the color of the car's interior) that would not have occurred to experts. The significant differences here are that first, the resultant range of criteria is oriented to how people *actually* solve the problem of buying new cars as opposed to how experts intuit that the problem *should* be solved, and second, that the range is organized and presented to the user in a process oriented manner (i.e., a problem solving time line).

Other evaluation comparisons might include asking another set of recent car buyers or current new car buyers to evaluate the usefulness of the various criteria and to see if they have concerns or questions which our system does not include (as opposed to a traditional



satisfaction measure). This would tell us not only where our system is incomplete or does not match what users need, but it will also give the system designers specific data on what has to be changed and how it needs to be changed. If we had a computerized format for our system, which would be organized in a manner very similar to our printed format, we could also test various network, computer platform and/or computer locations variables, all within the context of the usefulness vis-a-vis the actual task rather than testing user performance against an arbitrary expert-derived standard. Further, the evaluation criteria are derived from the task itself and its associated cognitive behaviors (e.g., reduction of uncertainty) rather than inherited from behaviors that are assumed to be relevant (e.g., ability to recall content, time on task, etc.).

## **A COMPARISON OF EVALUATION STRATEGIES AND CRITERIA**

Criteria used for evaluation of systems can be divided into two major categories: efficiency and effectiveness. Efficiency criteria are the best known in the area of formal evaluation and are accordingly readily recognized. Efficiency criteria deal with saving time and money, and with correspondence with an external standard. Examples would include:

- **coverage** - used here to compare the ranges of criteria between the expert-driven and the application-based approaches to our hypothetical DSS for car buyers. The contrast between the two examples illustrates differences in the knowledge acquisition procedures, in the validity of the resultant ranges, and in the ability of the ranges to meet the needs of the end users;
- **user demography/personality variables** - used here to explain why certain types of users do not find the system useful in the first DSS approach and to assure coverage of the problem solving behaviors and criteria through sampling procedures in the second;
- **accuracy** - used in the first DSS example to compare what experts think users *should* do as opposed to what the users *actually* do whereas in the second example, the assumption that there was no "correct" way to solve the problem (which is the case with the majority of the emerging automated applications) transformed this criterion into a "goodness of fit" between the aggregated user perceptions and the system model; and
- **time considerations** - usually operationalized as the time required for a user to accomplish a particular task associated with the intent of the system, they are reflected here as the time users spend accessing the DSS in the first example. In the second example, time is seen as a major task model organization logic,

taking advantage of how people experience the problem, how they tend to talk about it, etc.

Effectiveness criteria reflect a relatively recent concern in system and technology evaluation, primarily because effectiveness is oriented towards the system user while efficiency criteria have been oriented towards the system designers and/or the organization within which the user works. Given the surge in interest in "user friendliness" in the last few years, it is natural that effectiveness criteria are often seen as more useful. They tend to address the quality of the system in matching to the expectations and needs of the user. If carefully conceptualized and operationalized, they can also give the system designer significant insight to system structuring. Examples used here include:

- **satisfaction** - used in the first DSS example as either a binary measure (i.e., the user evaluates the system and/or system components as either satisfactory or not satisfactory) or, if pursued qualitatively, can indicate problems with the knowledge base, system implementation, etc. but no specific system improvement data. The analog to satisfaction for the second example would be to have car buyers use and comment on the system in terms of its completeness and utility which provide system designers with the "locations" of problems (i.e., where in the task model they are) as well as some insight into how the system has to change to be more useful;
- **system/user mismatch** - related to the use of the notion of accuracy in the effectiveness criteria above, this kind of criterion looks for differences in the way that systems are implemented (e.g., appropriate technology evaluations), the content of the system (e.g., differences in user perceived car buying criteria and those extant in the system), and the structuring of the system features (e.g., time ordered versus arbitrary organization);
- **usefulness/usability** - this criterion addresses concerns like the amount of training required (i.e., the training overhead) and ease of use from the perspective of the users, the ability of the user to "navigate" through the system (see Newby 1991 for a description of this specific criterion) in terms of the user's perception of where s/he is relative to the actual problem being solved and whether or not the user has a sense of where to go next. Although this criterion is being used more often in more conventional system evaluation, it focuses more on expert- or system-derived concerns rather than user perceptions; and
- **information needs** - associated with the use of any system is the information needed to use the system as well as the information (particularly in a DSS) needed for the user to maximize his/her benefits in solving the problem. In most system designs, the information functions are secondary and not conceptually

linked to the system content and structure.

This contrast of efficiency and effectiveness criteria is not meant to be an exhaustive comparison of these two categories, rather, it is intended to give the reader a sense that the application-based utility approach to system design and evaluation promises to add some valuable insight into the design of the increasingly complex applications now being developed. This in turn will provide client organizations (e.g., the government, private sector businesses, etc.) the basis upon which to make decisions about specific technological innovations like display technologies as well as system design features *in terms of their specific problem solving needs*. It is this latter capability that promises to be the most valuable in improving the development of efficient and effective systems.

The discussion above was intended to lay the groundwork for an examination of the applicability of this approach for the emerging technology of virtual reality (VR). What follows is a description of some of the features of the technology, followed by a description of its potential domains of use.

### **CUTTING EDGE SYSTEM TECHNOLOGY: VIRTUAL REALITY**

Virtual reality is a term given to the extension of the quality of immersion (i.e., providing the human senses with the illusion of existing in an electronic place) of the older but still recent technology called multimedia. This means that one or more of the human senses (seeing, hearing, touching, smelling, tasting) is presented with a stimulus that appears to be "real" and present (i.e., in three/four dimensions) rather than electronically generated through the presentation of visual, aural, tactile, etc. clues. Although the three levels or types of VR presented here are oversimplified and focus on distinctions between display and pointing devices, they do help to demonstrate a need for a different approach to evaluation.

There are (roughly) three levels or types of technology used in current applications of VR. The first level, which I call "low-tech" uses readily available, existing computer and electronic media technology (e.g., large 2-D screens, stereo or up to seven channel sound, video cameras, microphones, etc.) to provide an approximation of immersion. Large screen TV's with surround sound and cinemax theaters are examples of this level of technology. Particularly in the sound area, there are significant developments in the technology that have existing and expanding consumer markets (e.g., home theaters with Dolby Pro Logic

sound generated by high resolution laser disc players). By including computers with this kind of technology, VR designers have been able to suggest a wide variety of applications for this level of technology to include exercise applications where the user jogs or rides a bicycle through exotic landscapes with their attendant sounds. Existing non-media equipment (e.g., the stationary exercise bicycle or a treadmill) is used to provide tactile feedback in some applications. Some of the earlier (and still used) training and simulation applications which used several synchronized screens to simulate the various views out of an airplane cockpit and hydraulic platforms to provide tactile feedback belong in this category. Other application examples include interactive entertainment systems like games and music creation applications. Also at this level would be included some of the remote sensing and robotic applications on conventional 2-D display devices that use shading and perspective to give the illusion of 3- or 4-D images like architectural CAD applications. Finally, many data visualization and imaging applications could be described as this kind of VR.

The second level or type involves the use of special display, user tracking, pointing and tactile feedback technology that, for the most part, has been developed especially for VR. The current state-of-the-art example would be a set of "eyephones" like those developed by VPL that have one LCD television screen for each eye, the images for which are computer processed and delivered stereoscopically. This device is attached to the computer via an umbilical cord which either trails behind the user or is linked to a boom that keeps its weight and cumbersomeness at a minimum. Also attached to this device are a special tracking device (usually magnetic although sometimes ultrasonic) that tells the computer where the user is in relation to the electronic space and a set of headphones for sound. Although binaural sound cues for simulating 3-D sound is technically possible, this embellishment is seldom incorporated into VR setups; simple 2 channel stereo sound is more common. Also included in this type of VR is some kind of pointing device, with or without tactile feedback. One of the most common devices is called a "data glove" which incorporates another tracking device to tell the computer where the user's hand is and some fiber optic devices that tell the computer the positions of the fingers (i.e., whether the user is pointing a finger or making a fist, etc.). Some applications at this level also include non-media equipment to provide more holistic tactile feedback like an architectural design system at the University of North Carolina at Chapel Hill that uses a treadmill with

handlebars in place of the data glove and tracker combination to allow the user to navigate through electronic buildings.

Some of the existing applications of this level of technology include medical imaging for planning delicate surgical operations and for focusing and aiming radiation beams for cancer treatment. There are also a variety of immersion entertainment applications which are still in the development stages. One example of the latter is a project reportedly underway in Japan to create a VR amusement park where participants would feel like they were riding a rollercoaster. Although significant and important advances are being made, many severe limitations to applications in this category of VR exist aside from the cost of most of these devices, which is astronomical. These would include the need for an umbilical cord to link the user to the computer system, the less-than-satisfactory resolution of the display devices (e.g., the VPL eyephone cannot display easily readable text), the lack of processing power to create/process and present real-time images, realistic tactile feedback mechanisms, high speed networks to deliver data from distributed databases and/or to deliver signals to multiple users who are not co-located, and database and network management procedures for synchronizing and integrating the various data types.

The last type of VR represents what are called "volumetric" display technologies with more naturalistic human-computer interaction capabilities. In the display area, volumetric displays tend to be generated using laser technology, either holographic or where a computer controls a laser which is focused on pulsing mirrors or rotating helixes creating a 3- or 4-D image that requires no paraphernalia physically attached to the user. Further, more than one user can see the displayed image. Human-computer interaction in this type of VR also frees the user from umbilical connections to the computer with voice recognition interfaces and low-powered laser pointing devices aimed at a light sensitive display surface (for menu selections, etc.). Unlike the first two types of VR, this type is making use of very different technologies, many of which have not been adopted or derived from existing technological applications. Suggested applications here are in multi-user contexts or for elaborate/hazardous remote control situations (e.g., NASA's full-parallax holographic display system for remote operations). One possible example would be a multi-user command, control, communication and intelligence (C<sup>3</sup>I) system with subsystems that deal with user training, planning, crisis management as well as on going tactical command. The limitations of this type of VR are included in the second level but

also include the nascent state of voice recognition technology, providing color to holographic and laser derived images, and, more obviously than in the other levels, system design capability because of the complexity of envisioned application possibilities.

Conceptually, there are three sources of images for the visual aspects of VR: digital and/or analog representations direct from the physical environment (e.g., stereo camera images), images derived from specific data points and streams (e.g., much of the work done in the data visualization area), and artificial data (e.g., animations). Each of these sources will ultimately play a significant role in creating usable VR systems and each of them represent significant and somewhat unique technical problems before they become reliable for envisioned VR applications.

Most of the work currently being done in the VR area focuses on the development of the technology itself - display devices, pointing devices, tactile feedback, image processing capability and strategy, data visualization, etc. Almost all of this effort is directed towards VR emulations of actual physical reality. It is *very* important to note that VR presents another possibility that is equally compelling and perhaps ultimately much more valuable.

Human beings have been able to process about 300 characters per second (on the average) since the development of writing. None of the technological innovations in information production, dissemination and display to date have improved on that processing rate - people still read about 300 characters per second. However, reading is a unimodal sensory capability. Researchers in communication, perception, etc. know that multimodal capacities are more accurate and, depending on the nature of the sensory input data, capable of exceeding the bandwidth of any single sensory channel. For example, human beings tend to "lead" with their eyes when trying to identify the direction a particular sound is coming from but very quickly (and often) correlate auditory clues with visual clues. Together, these two sensory modes are more accurate than either one of them alone.

If we take this understanding and add it to the potential bandwidth capability of VR technology without that bandwidth being devoted to maintaining a physical reality analog, what we have is a potential for organizing and representing data and data streams such that users can exceed the 300 character per second limit. The only example of this that I have heard about is the rumored, on again - off again project (currently rumored to be on again)

involving IBM and NYNEX (among others) where the goal is to transform and display all the information and data that a Wall Street financial analyst needs to make an intelligent purchasing decision on a particular stock. The idea is that by transforming the existing extensive alphanumeric data streams currently seen as useful in making such a decision into multimodal representations in a four dimensional volume, the analyst might be able to make sense out of these diverse data streams faster and with more accuracy than is currently possible. To some researchers, myself included, this possibility promises ultimately to be much more valuable than the current emphasis on physical reality representations.

## **PUTTING VR AND APPLICATION-BASED UTILITY EVALUATION TOGETHER**

Given that current system design and system evaluation are based upon technology-driven and expert-driven approaches (see Taylor 1986 for a more complete discussion of this issue) and, given that the current VR development emphasis is following in this trend, it seems unlikely that the exploitation of VR for increasing the human-computer interaction bandwidth will proceed as quickly or as productively as it might by a more coherent inclusion of the cognitive behaviors associated with the problem that the system is intended to solve (i.e., a more comprehensive accommodation for the application-based utility of the various system hardware and software features). While I might argue that the exploitation of VR technology for the current physical space analog applications would also be more efficient and effective by the same argument, it is the second, and largely ignored, application area that is more compelling.

The proposed application-based utility approach provides several conceptual, methodological and analytic advantages over existing system design and evaluation approaches which, when used in conjunction with the existing approaches promises to facilitate the timely development of more usable systems in this age of increasing complexity. To wit, the proposed approach:

- focuses much more coherently on the problem to be solved;
- systematically employs the perceptions of system users in knowledge base development and in system configuration;
- provides an application/task model that can be used to evaluate the appropriateness of technological system features (e.g., when does increased

resolution actually help the user do what s/he needs to do);

- provides evaluation criteria and methods that address gaps in more conventional evaluation approaches;
- draws from human-computer interaction perspectives that focus directly on the cognitive communication behaviors in the context of a particular problem;
- can provide for a significant reduction in training overhead; and
- can be scaled up to provide design insight for massive, complex systems.

## NOTES

1. If indeed it was ever straightforward. For example, one can look around any organization and see countless examples of \$10,000.00 technical solutions to 10-cent problems like work stations with high resolution graphics displays being used exclusively for word processing.
2. See Newby, Nilan & Duvall 1991 for a discussion of the individual differences approach to system design.

## REFERENCES

- Agarwal, R. Knowledge Acquisition for Business Expert Systems: A Decision-Centric Model and Its Empirical Validation. An unpublished Ph.D. Dissertation, Syracuse University, School of Management, 1988.
- Dervin, B. and Nilan, M. S. "Information Needs and Uses," in M. E. Williams (Ed.), Annual Review of Information Science and Technology, Vol. 21, 3-33, Knowledge Industry Publications, White Plains, NY: 1986.
- Hert, C. A. and Nilan, M. S. "User-Based Information Retrieval System Interface Evaluation: An Examination of an On-line Public Access Catalog." Paper to be presented at the 54th Annual Meeting of the American Society for Information Science, Washington, D.C., October, 1991. Published in Proceedings of the American Society for Information Science, Volume 28, 1991.
- Kanaan, R. Control Strategies for a Knowledge-Based Purchasing Decision Support System. An unpublished Ph.D. Dissertation currently under revision, Syracuse University, School of Management, 1991.



Newby, G. B. "Navigation: A Fundamental Concept for Information Systems with Implications for Information Retrieval." Paper to be presented at the 54th Annual Meeting of the American Society for Information Science, Washington, D.C., October, 1991. Published in Proceedings of the American Society for Information Science, Volume 28, 1991.

Newby, G. B., Nilan, M. S. and Duvall, L. M. "Towards a Reassessment of Individual Differences for Information Systems: The Power of User-Based Situational Predictors." Paper to be presented at the 54th Annual Meeting of the American Society for Information Science, Washington, D.C., October, 1991. Published in Proceedings of the American Society for Information Science, Volume 28, 1991.

Nilan, M. S. "User-Based Information/Communication Systems: Using Sense-Making to Create a User Model for a Desktop Publishing Help and Training System." Peer reviewed chapter accepted for B. Dervin (Ed.), Methodology Between the Cracks: Theory and Method for the Qualitative and Quantitative Study of Human Communication. Norwood, NJ: Ablex Publishing Company, May 1992. Copies of the original draft are available from the author.

Taylor, R. S. Value-added Processes in Information Systems. Norwood, NJ: Ablex Publishing Corp., 1986.

# FINAL REPORT

## OPTICAL FIBER AMPLIFIERS AND OSCILLATORS

Sponsored By

AFOSR Summer Faculty Research Program

At

The Air Force Photonics Center  
Griffiss Air Force Base

Kenneth J. Teegarden  
Institute of Optics  
University of Rochester  
Rochester, New York

Salahuddin Qazi  
Department of Electrical Engineering  
Institute of Technology  
State University of New York  
Utica, New York

Summer, 1991

# OPTICAL FIBER AMPLIFIERS AND OSCILLATORS

Kenneth J. Teegarden  
Salahuddin Qazi

## I. Introduction

Optical amplifiers and oscillators based on fiber waveguides doped with luminescent impurities have been studied for many years. Recently the advantages of in line optical amplifiers for fiber communication systems have been recognized and this has led to intense efforts to develop amplifiers for the wavelengths where commercial monomode optical fibers have the lowest loss and can be designed to have zero dispersion, namely at about 1.30 and 1.50 microns. The most successful amplifier to date is one using the rare earth erbium which luminesces at 1.55 microns. Developmental models of erbium doped fiber amplifiers which achieve small signal gains of up to 30db and power amplification of 13dbm are currently on the market and are being evaluated in the field. These amplifiers are pumped with relatively expensive strained quantum well laser diodes operating at .980 or 1.48 microns. An increase in available pump power and a reduction in cost could be achieved if efficient operation at a pump wavelength of .800 microns was possible and the discovery of a glass composition which would eliminate excited state absorption of this pump wavelength would be an important breakthrough. Also, an increase in the optical bandwidth of erbium based amplifiers would permit the advantages of wavelength multiplexing to be more fully realized. This again is a problem whose solution depends on the development of glass compositions which widen the gain curve of erbium, even at the expense of maximum gain.

Most of the optical fiber embedded today operates at 1.3 microns and no practical amplifier based on silica fibers has as yet been developed for that wavelength although some promising systems based on neodymium in non silica glasses have been discovered. The search for a 1.30 micron fiber amplifier continues to be vigorously pursued in several laboratories with emphasis being placed on the discovery of new dopants for silica fibers with transitions at 1.3 microns.

Recently, erbium doped fibers have been employed as optical oscillators, or lasers, as well. In particular, they have been used to construct tunable, mode locked, traveling wave ring lasers that produce trains of 50 ps wide pulses with repetition rates of 6GHz. Such oscillators are proposed for use in very high band pass digital communications systems where they would replace solid state diode lasers as the primary signal source for all-fiber communications systems. The application of fiber ring lasers to communications systems is not as advanced as is the case for fiber amplifiers. No developmental models are currently available, indicating that the parameters for stable operation of the device have not as yet been established. It is

clear, however, that the same materials problems exist for the oscillator as exist for the amplifier. There is a need for dopants and glass compositions that increase tuning range and permit operation at other wavelengths, such as 1.3 microns.

The work described herein was carried out in the Air Force Photonics Center, Rome Laboratories, Griffiss Air Force Base, New York. Its objective was to establish in house expertise in the area of fiber based luminescent amplifiers and oscillators as a foundation for subsequent efforts to develop new devices and incorporate them into communications systems of interest to the Air Force. The specific project, undertaken during the summer of 1991, was the construction and characterization of a state of the art erbium doped fiber amplifier using commercially available components.

## II. Experimental

The experimental set-up used in most of the measurements is shown in figure 1. The erbium fiber used as the gain medium was supplied by Dr. W. J. Miniscalco, GTE Laboratories. It had a core diameter of 8.0 microns and an outer diameter of 100 microns. A commercial multiplexer (JDS FITELE Inc., model # WD915F-9S/N Y53) was used to combine an optical signal at 1.5275 microns with pump light from various sources. The single mode optical signal was generated by a distributed feedback laser in a commercial lightwave transmitter (Lasertron model # QLX-1000). Preliminary data using a frequency doubled Nd:YAG laser (ADLS Model # 300) operating at .530 microns as a pump source was obtained, but only data using a strained quantum well laser (Spectra Diode Labs Model # S9075) operating at .980 microns as pump is included in this report. Commercial single mode fiber patch cords designed to operate at 1.3 microns and terminated with FC/PC connectors were fusion spliced onto both ends of the amplifier fiber, and onto the input end and the signal output end of a wavelength demultiplexer (Gould, Inc. model # 980-1550-CXW-MX-02x02-01). Similar fibers with connectors were fusion spliced onto the output and signal input of the multiplexer, while the pump input end of this device was left bare, carefully cleaved and mounted in an xyz positioner. Light from the pump sources was collimated when necessary and focused onto the core of the multiplexer input with appropriate microscope objectives. Although the connectors used in this set up represented avoidable losses, they were used so that measurements of optical power could be made at various points in the fiber system. In a real amplifier they would have been replaced by fusion splices.

The overall gain of the amplifier and its emission spectrum were measured as a function of pump power and signal power for the pump wavelengths given above. The emission spectrum was determined using an optical spectrum analyzer. The overall gain of the system was measured in two ways: either the peak output power was measured on the spectrum analyzer for a measured peak input power, or the lightwave transmitter was amplitude modulated with a 33 MHz sinusoidal RF signal and the resulting AC input and output signals measured with a commercial lightwave receiver (Lasertron model # GRX-700) and an oscilloscope. In both cases the input signal was

measured at the output end of the demultiplexer, thus losses in this device were excluded from the determination of overall gain. Measurements of pump power injected into the amplifier were made with a Newport power meter ( model # 835), again at the output end of the demultiplexer.

A determination of the gain of the doped fiber alone was made by separately measuring the losses at 1.55  $\mu\text{m}$  which occurred in the demultiplexer, connectors and fusion splices and correcting the overall gain for these losses. A Fotec power meter was used for these measurements.

### III. Results

#### A. Emission Spectrum

The emission spectrum of the erbium doped fiber pumped at .980 microns with 20 mw. of power is shown in figure 2. This spectrum is identical to data published in the literature to within the calibration of the spectrum analyzer used, except for a periodic ripple which is superimposed on the more slowly varying envelope of the fluorescence. The magnitude of this ripple and the shape of the emission curve were dependent on pump power and the exact way the pump radiation was launched. For example, if the end of the input fiber to the multiplexer was moved very slightly from the position corresponding to figure 2., the spectrum shown in figure 3a. was obtained. Here the ripple is much more pronounced. The ripple moved slowly in time relative to the emission band and this gave rise to periodic fluctuations in the gain of the amplifier.

If the connectors in the systems were loosened and held in the right position, even more drastic changes in the emission spectrum occurred. This is illustrated in figure 3b., where the emission at peak of the spectrum appears greatly enhanced as well as the ripple. In fact at maximum pump power, obvious lasing action could be induced as can be seen from figure 4. These effects most likely occurred because of feedback from reflections at coupler interfaces or fiber ends and could be minimized by careful alignment and tight couplers. It should be noted, however, that because they represent oscillations in the gain curve they cause instabilities in the gain of the amplifier. An improvement in amplifier operation would have occurred if optical isolators were included in the amplifier design, and this is planned in future work.

#### B. Amplifier Gain

The overall gain of an amplifier based on a 4.75 meter length of erbium fiber was measured using the first method described above. Figure 5 illustrates how this measurement was made. Figure 5a shows the signal laser power at the output of the multiplexer after attenuation with a variable optical attenuator and figure 5b gives the amplified output for a pump power of 20 milliwatts at .980 microns. Note that the peak of the gain curve for erbium lies at about 1.5303 microns so that the signal in this case did not coincide with the wavelength for maximum gain. The values of gain reported here are therefore not peak values but are about one half the gain expected at 1.5303

microns. The overall gain calculated from the ratio of the output to the input signal in this case is 9.0 or 9.5 db. The overall gain of a 4.75 meter long fiber measured in this way as a function of launched .980 micron pump power for an input signal power of 30 mw is shown in figure 6.

The transmission of the 4.75 meter fiber at .980 microns as a function of launched pump power ( power at the out put of the multiplexer) is shown in figure 8. Clear signs of saturation are apparent in this figure. That is, the fiber was more transparent at a launched power of 20 mw. than at a power of 1.0 mw. The data shown in figure 7. were used to correct the data in figure 6. and generate a plot of the overall gain of the amplifier as a function of absorbed pump power rather than launched pump power. The result is shown in figure 8.

The data shown in figures 6. and 7. indicate that a higher gain would be obtained by using a fiber of longer than 4.75 meters. At 20 mw of pump power, 10mw was transmitted and this 10 mw would produce a gain of about 4.0 in a fiber of comparable length. Therefore, an additional three meters of erbium doped fiber was fusion spliced to FC/PC connectors and coupled to the original 4.75 meters of fiber. Measurements of overall gain were made on this 7.75 meter fiber using the second method described above in which the signal laser was amplitude modulated with a 33MHz signal. The results of these measurements are shown in figure 9, where overall gain is given as a function of pump power and relative signal power for an input signal modulation of 1.0 mV. peak to peak. A strong dependence on launched pump power and considerable saturation at maximum signal power are illustrated by this data. As predicted, this greater length of fiber showed a higher gain even though the additional couplers and fusion splices introduced additional losses.

Finally, the gain of the doped fiber itself was estimated by correcting the maximum overall gain obtained above for losses incurred in the FC/PC bulkhead connectors and the two fusion splices between the standard size fiber patch cords and the erbium fiber. Losses which occurred in the fusion splices because of mismatch between core diameters were by far the largest and were found to be 2.7 db for the two splices in series. The measured losses in the connectors was about .05 db per coupler on the average. Since there were three couplers involved, the total loss amounted to about 2.9 db. Thus the peak fiber gain must have been about 14 db. at a launched pump power of 20 mw. and a pump wavelength of .980 microns. It should be noted however, that the fiber core mismatch in the fusion splices occurred because the core of the doped fiber is made smaller than the standard size of 10 microns in order to increase pump power density in the amplifying fiber. Thus a major loss in this system is inherent in the way the fiber is designed. Also, in an operational system losses in both multiplexer and demultiplexer must be included in a specification of overall amplifier gain. These were found to total 3.2 db at the signal wave length of 1.5275 microns. Finally, we noted above that the wavelength of the signal did not match the peak of the erbium gain curve. In fact the peak gain of the fiber was about twice that of the gain measured at the signal wavelength, that is 50 or 17 db.

#### IV Summary

This project illustrated some of the basic problems to be considered in the construction of an erbium fiber based optical amplifier. The gain of the amplifier was strongly dependent on the launched pump power and the efficiency with which pump radiation is injected into the fiber is therefore of critical importance. A pig tailed pump laser might have increased the pump power available in the present investigation. The dependence of the gain on fiber length is also an important factor in amplifier design, but was only briefly investigated in this work. The major losses which placed a limitation on the overall gain of the amplifier were found to lie in the core mismatch between doped fiber and standard fiber. One way to avoid this problem would be to increase the core diameter of the doped fiber to 10 microns and offset the reduction of pump power density by increasing launched pump power. Since pump lasers are continually being improved, this is probably not an unreasonable option. Losses in the wavelength multiplexer and demultiplexer used were significant, but may be reduced in future models of these devices. Finally, if higher gains are to be obtained in the amplifier constructed in this project, inclusion of optical isolators to reduce internal feedback and resultant instabilities is of critical importance.

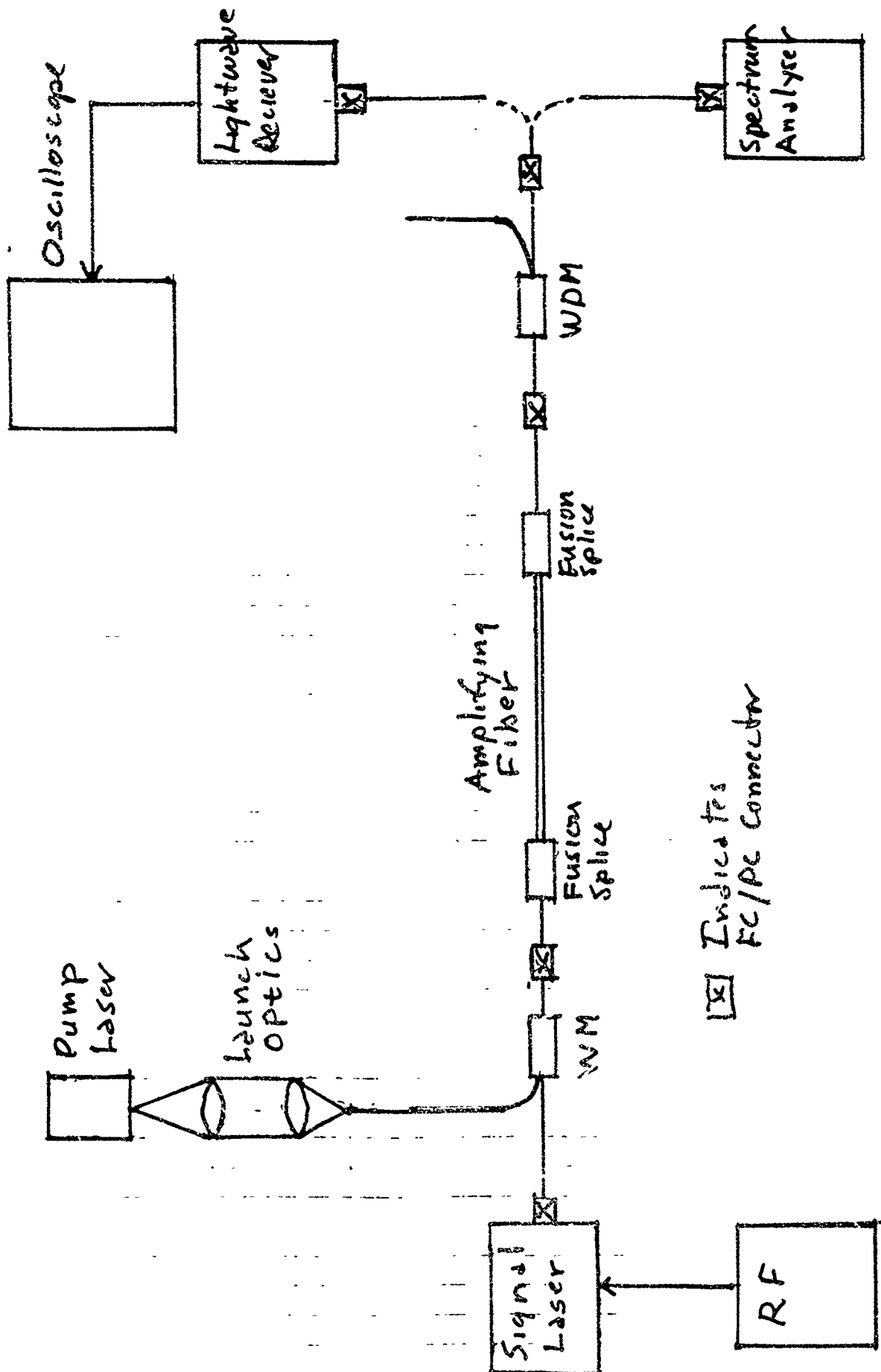


Figure 1.



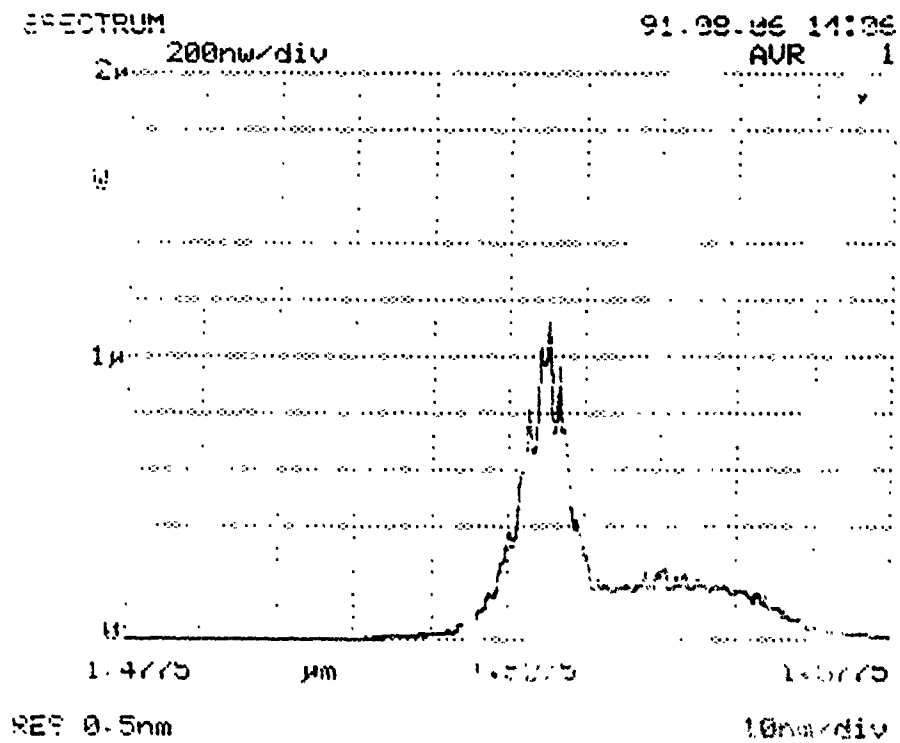


Figure 2

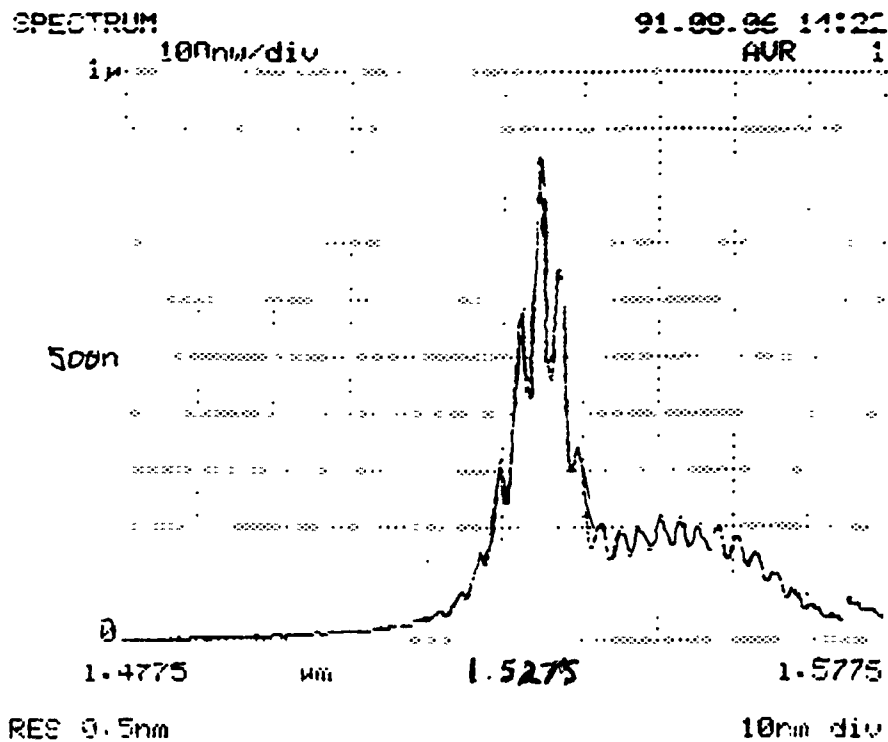


Figure 3a

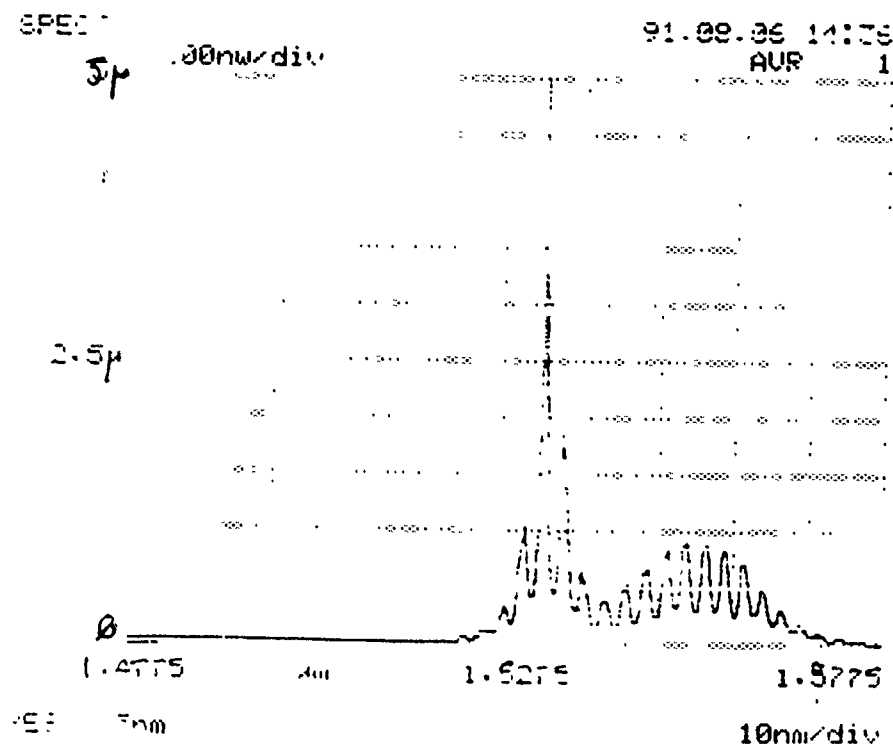


Figure 3b

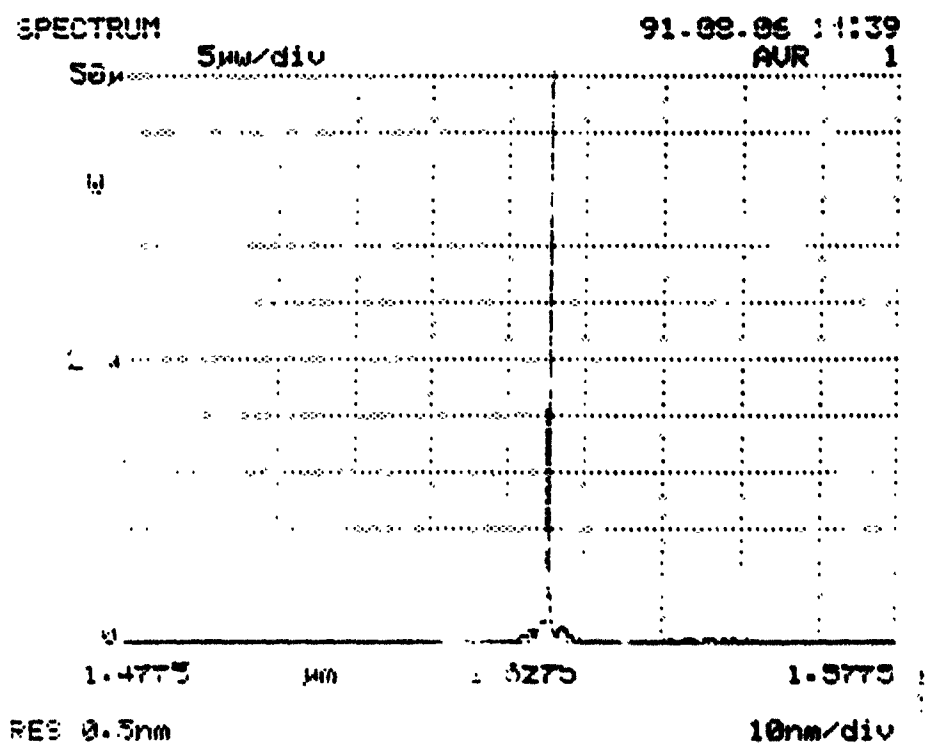


Figure 4

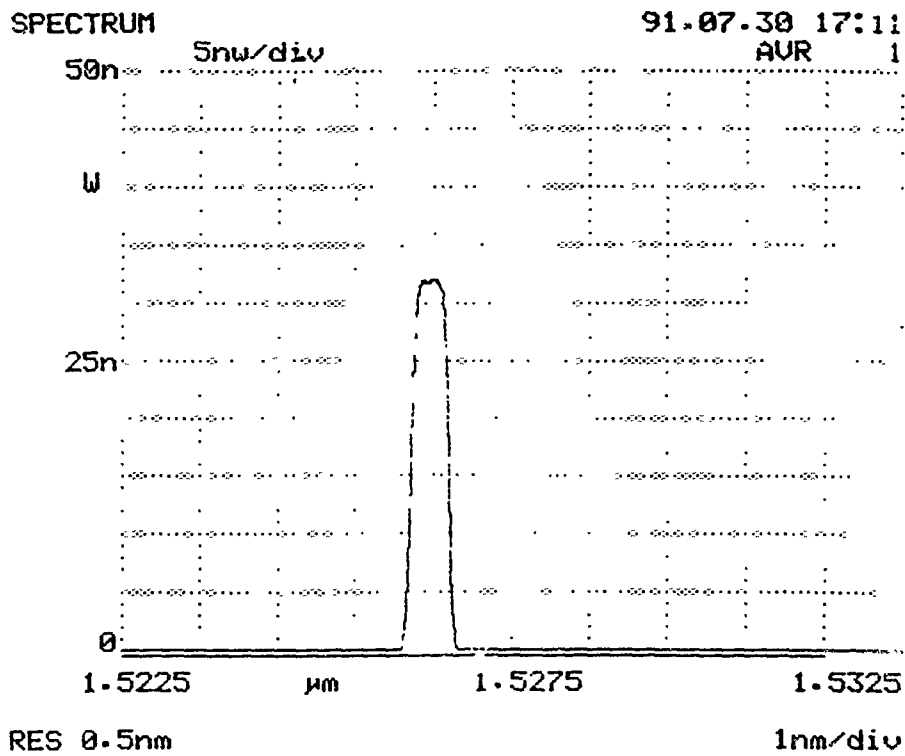


Figure 5a

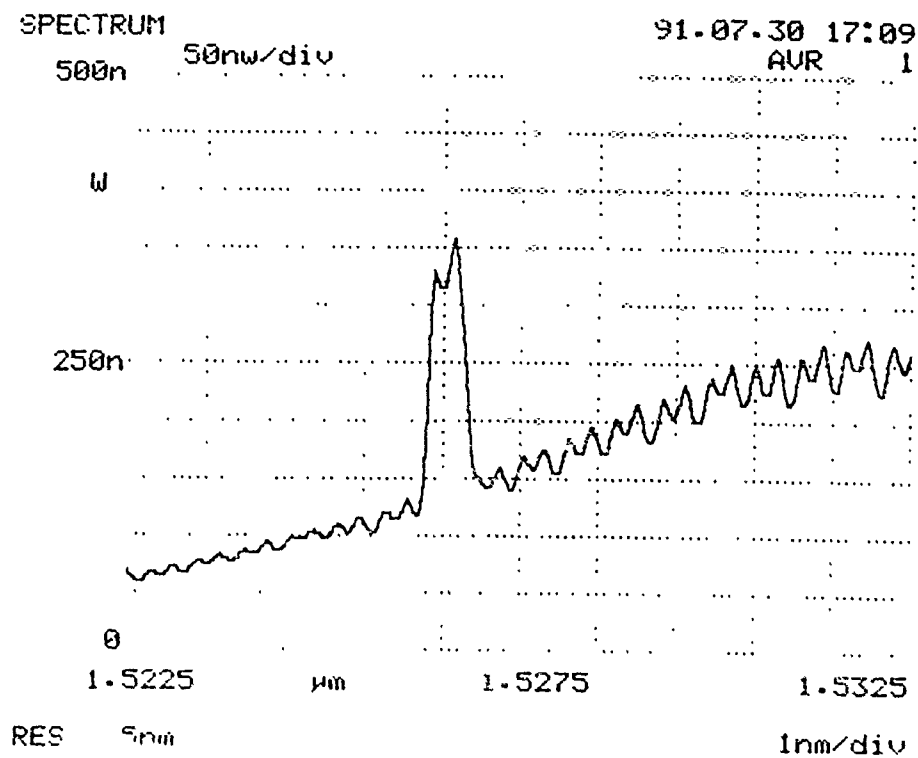


Figure 5b

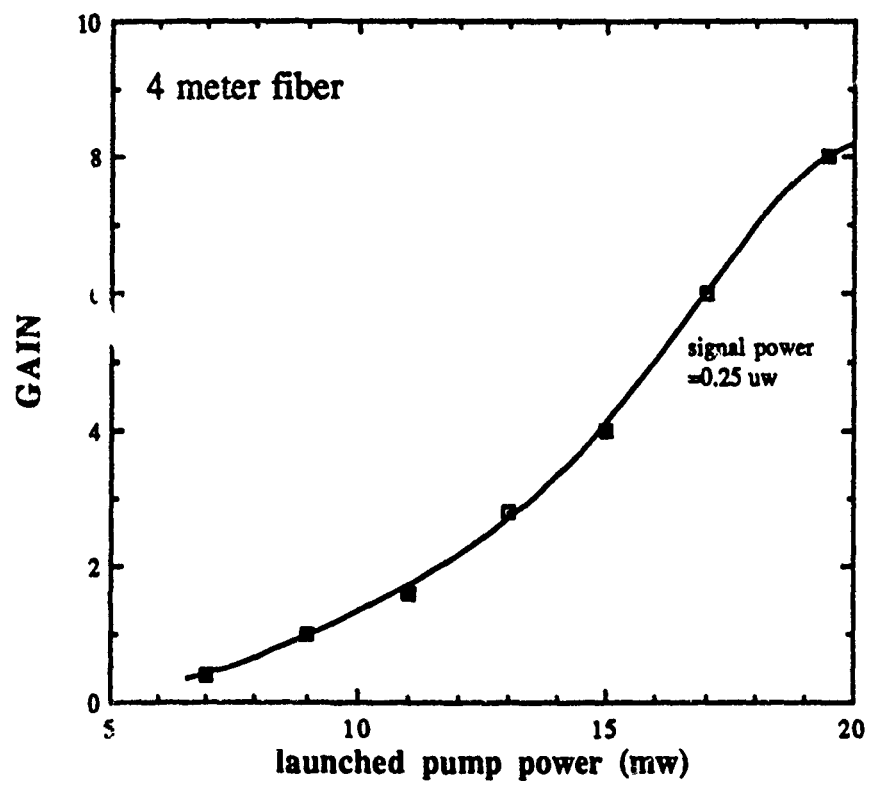


Figure 6

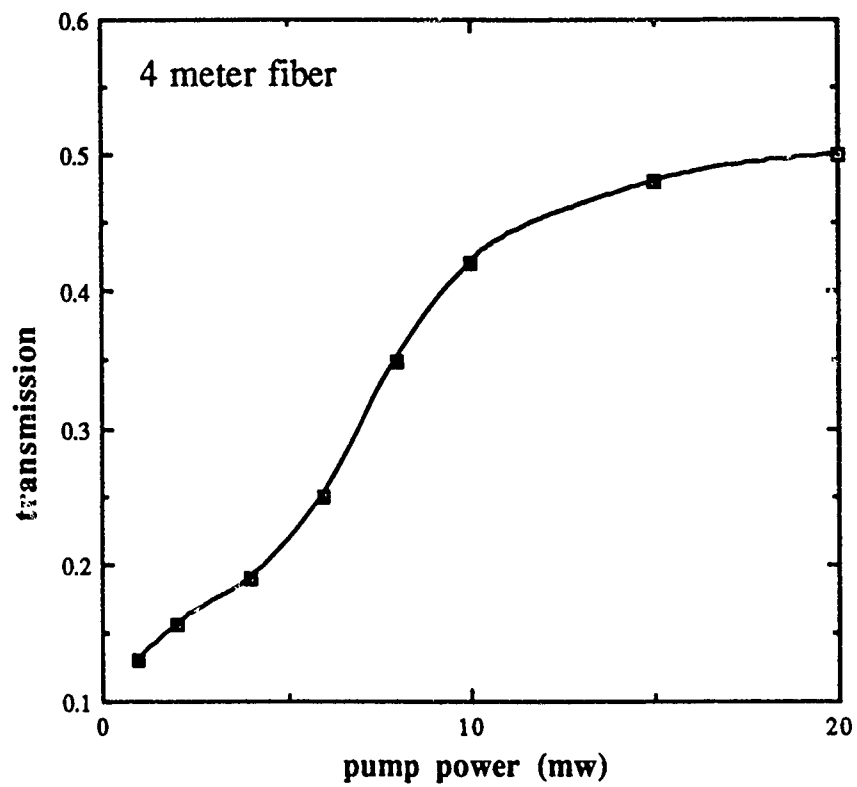


Figure 7

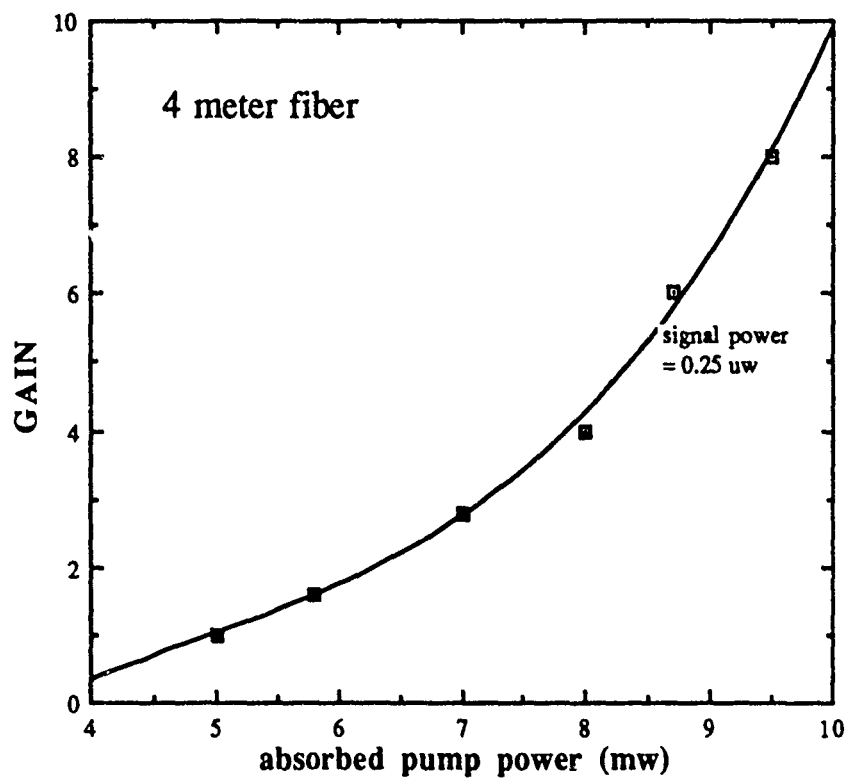


Figure 8

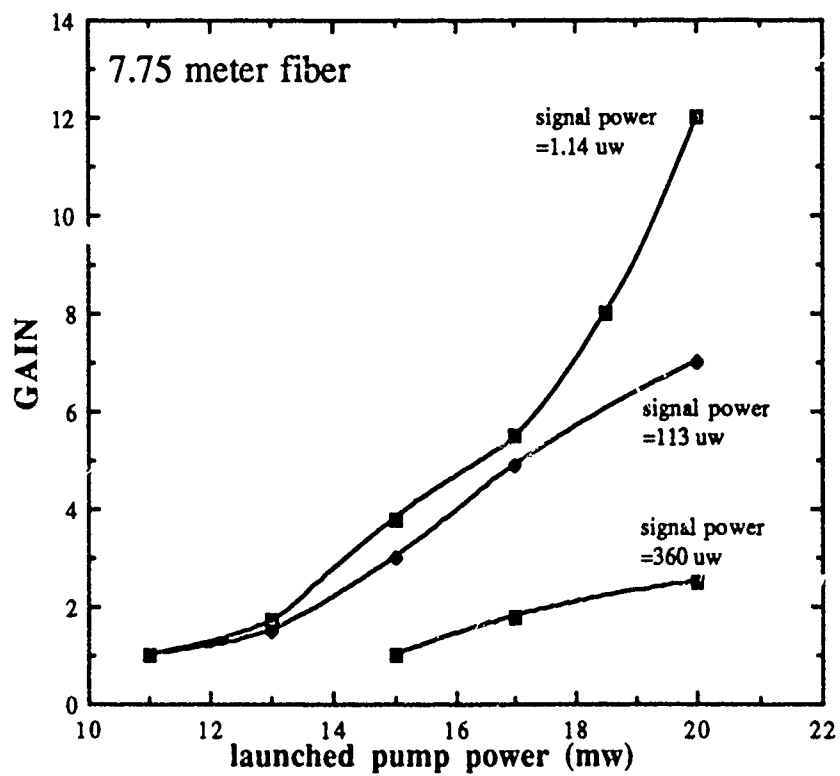


Figure 9



Collecting Data for Markov Models of Error  
Patterns on Data Communications Links

Wayne D. Smith

ABSTRACT: Rome Laboratory is currently in the process of procuring Error Injector Units (EIU) to model the error behavior of data communications channels. A major component of this effort deals with finding Markov models that will simulate the behavior of these channels. The objectives of this effort were to continue the research initiated in 1990 to find Markov tables to load into the EIUs. A portion of the research involved searching current literature for Markov tables or error gap data. Alternative approaches to obtaining the desired Markov tables were also considered. This included research into the use of simulation models to produce Markov tables, and some further work on the concept of "heuristic error models." The most substantial effort was devoted to the design of a data collection experiment that could be used at Rome Laboratory to produce the needed data. This effort included the design of a microcomputer interface that would permit a microcomputer to monitor a "Fireberd" bit error rate tester and record the error statistics for analysis and conversion to a Markov table.

## I. INTRODUCTION:

Rome Laboratory is in the process of procuring specialized computer hardware that can be used to model the error behavior of a number of different data communications channels [1]. The hardware can be placed in line in any communications link, and can be used to simulate the occurrence of errors on that link.

As each bit is received at the EIU input, the logic of the unit determines whether or not an error is to be generated for this bit. If an error is to be generated, the incoming bit is exclusive ored (EXOR) with a one bit produced from within the EIU board. This operation results in an inversion of the incoming bit at the output port of the EIU.

The EIUs are based on a Markov chain that models the behavior of the data communication links. The model is internally represented with an  $n$  state ( $n \leq 8$ ) Markov model, and transition between states is controlled by an  $8 \times 8$  matrix, called the Markov table. A random number generated within the EIU circuitry determines the next state of the model.

The Markov tables can be down loaded to the EIU from a SUN controller, or entered directly from the EIU microcomputer keyboard. Once the table has been loaded, the EIU will model the error behavior of the link being tested as determined by the Markov model. A major component of the

EIU project is concerned with finding Markov table values that will produce an error bit string that will mimic the behavior of various data communications links.

It is well known that data errors tend to occur in clusters or "bursts" that are divided by relatively long periods without errors, called gaps. Any model of data errors must take this "bursty" nature of real errors into consideration. In order to characterize data errors in general, it is convenient to define a simple data communications channel model as shown in Figure 1.

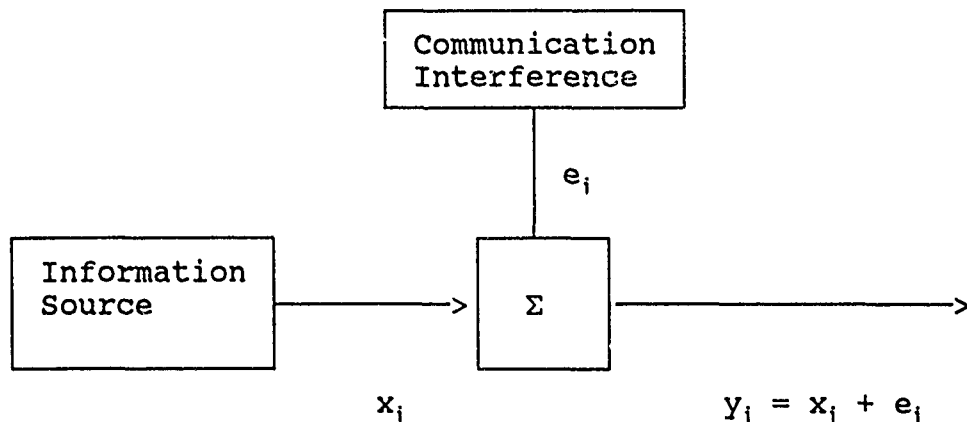


Figure 1: Data Communications Model

In the model,  $\{x\}$  represents the bit string of the transmitted data. The channel corrupts the incoming data with interference and produces the output data string  $\{y\}$ . The string  $\{e\}$  is the error string. A one in the string  $e$  represents an error, while a zero represents the absence of an error. The "+" operator represents modulo 2 addition or an exclusive or operation.

The most convenient way to characterize the error distribution involves specifying the gap distribution of the data. A gap is the opposite of an error burst, and is represented in the error string,  $\{e\}$ , as a sequence of zeroes in the string. The gap distribution is a measure of the number of times that a gap of length  $m$  occurs in a particular error string. By definition, a gap is preceded by an error, so the gap error probability is written as:  $P(0^m|1)$ .

Gilbert [3] showed that the error performance of a data link can be characterized by a two state Markov model, where one state is the one that may generate an error (B), and the other state is error free (G). Discrete probabilities are associated with the generation of an error in the B state, and also with the likelihood of a transfer from one state to the other.

Fritchman [4] extended the Gilbert model to an  $N$  state model, with  $k$  error free states, and  $N-k$  error states. He also showed that an  $N$  state model with a single error state, and  $N-1$  non-error states is sufficient to represent the data distribution from a data communications link. State transitions in the model are governed by a Markov table.

In striving to model the error performance of a data communications link, the goal is to discover a Markov model that will produce an error sequence with the same distribution as the observed error data. A number of

researchers have collected data on various data communication link error characteristics, and have used this data to define the structure of Markov chains that would reproduce the observed error distribution [4, 5, 6, 7, 8]. Varshney [9] has collected much of this data in an RADC Technical Report.

## II. DISCUSSION OF THE PROBLEM:

The objectives of the research effort in the summer of 1991 were to continue the work initiated in 1990 to find the Markov tables needed to permit the EIUs to model the characteristics of as many different data channels as possible. This effort included additional literature review for any previous research that had produced Markov models that were suitable for use with the EIUs. Some effort was also devoted to a literature search for error data that was not already in Markov table form, but that was suitable for conversion to a Markov model.

As an adjunct to the literature review, alternative methods for obtaining the Markov tables were investigated. This included some work on the heuristic models started last summer, as well as the potential use of simulation models of data communications links to provide gap data that could be reduced to the Markov table format.

With the realization that these methods would probably not produce the quantity and type of error data needed for

the EIUs, the major effort was devoted to finalizing the design of an experiment to collect the needed data at Rome Laboratory. This design included the development of experiment procedures, the design of a computer interface between a microcomputer and a bit error rate tester, and the initial specification for the software needed in the microcomputer to collect the needed data.

Finally, the facilities at the University of Mississippi that are used to reduce the gap data to Markov tables was tested, and the access procedures from Rome Laboratory were developed. The reduction of gap data from a Markov table computer simulation program was used as a test case.

### III. RESULTS:

The literature review was unsuccessful in locating any additional Markov tables for use in the EIUs. Although the search continues, it appears unlikely that any new models will be discovered.

In the study of the use of simulators of various types to produce Markov tables, two papers devoted to this subject were discovered. One of these involved an analog simulation of the Rayleigh fading model [11]. The second was a digital computer simulation of the same model with similar parameters [12]. Together, these two papers provided

several additional Markov tables, albeit several levels removed from real world channel error performance.

In spite of almost one year of extensive literature review, including computer data base searches, to date very little data on communication channel errors has been uncovered. The data that has been found is relatively old, and represents equipment that is no longer applicable the data communications needs of the U.S. Air Force.

One reason for the lack of success in this search deals with the format of error data required by the EIU project. The EIU is a far more sophisticated device than any that have been used to emulate data errors in the past. The EIU emulates not only the error rate, but the distribution of errors within bursts. This produces models with much higher fidelity than other types that tend to use an independent error assumption.

This higher fidelity is not without cost, however. In order to construct a Markov table that models the communication errors requires complete information about the error distribution. This means that the error data must be collected in detail. Just counting the number of bits received and the number of errors is not adequate. Data with this level of detail is more difficult to collect, and requires more storage capability than just collecting error rates. Hence, the amount of information suitable for conversion to Markov tables is in rather short supply.

As a result of the lack of success in finding either Markov tables or appropriate data for use in producing such tables, significant effort was invested in the design of an experiment to collect the needed data at RL. The first draft of a proposed design was presented in a "white paper" during a visit in March of 1991. As this paper met with some acceptance, additional effort was devoted to the refining of the design of an experiment to collect the needed data on-site at Rome Laboratory.

To perform the required data collection, it is necessary to set up a transmitter at one location and a receiver at another. The transmitter would transmit a series of random data bits over the communications channel. At the other end, the receiver would compare the data bits received with the bits actually transmitted, and record the occurrence of errors. For the purposes of the initial experiment, no error detection or correction techniques would be used with the data.

At the receiving end, the receiver would have to generate a sequence of random bits identical to that used by the transmitter. The receiver must compare these bits to the incoming bits and detect errors within the data stream. The receiver must also record the errors that are detected. This task will require a mass storage device for holding the collected data.



Figure 2 shows a block diagram of the equipment needed to set up the data collection experiment. The blocks on the left represent the transmitter and those on the right are the receiver.

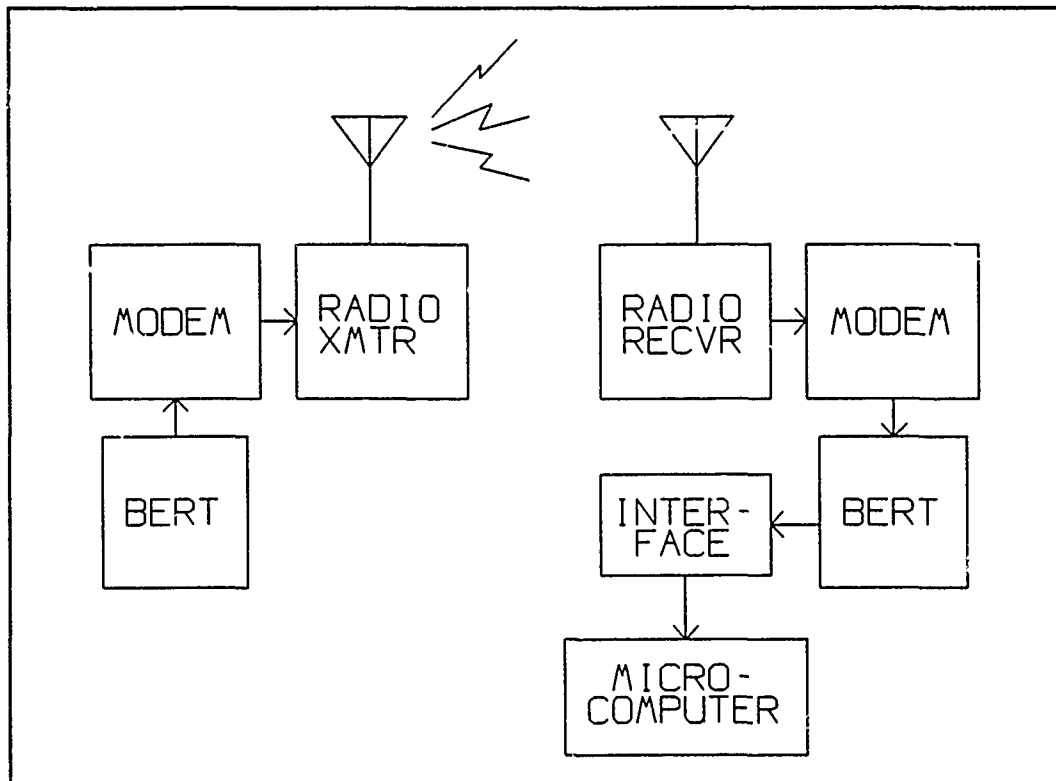


Figure 2: Data Collection Experiment Hardware

An essential piece of equipment for use in this experiment is the Fireberd 6000 (or equivalent) Bit Error Rate Tester (BERT). This unit can be programmed generate several different types of random bit sequences for use in the data collection experiment. By using two such units, one at the transmitter, and one at the receiver, the generation and comparison of random data bits becomes relatively simple. When first initiated, the receiving BERT will synchronize itself and lock on to the bit sequence from

the transmitter. It will remain in synchronization until a loss of signal or extremely long error burst results on a loss of synchronization.

By connecting the BERT units to a transmitter and receiver, some error statistics pertaining to the data channel can be collected directly by the BERT. These statistics include error counts, bit error rates, block error rates, etc. Unfortunately, the statistics collected by the BERT are not sufficiently detailed for the construction of Markov tables.

In order to collect the error gap data needed for the construction of the Markov tables, a computer must be included in the system. The BERT unit provides a number output signals that will permit the interfacing of the BERT to a microcomputer dedicated to collecting the error gap data. These signals include a loss of synchronization (LOS) signal, a recovered clock (CK) signal and an error (ERR) bit signal.

In essence, the error collection system would work by reading these status bits from the BERT. Each time that the CK signal is received (without ERR), the gap size count is increased by one. When an ERR signal is received along with the CK signal, the gap size count is stopped, and a value representing the number of times a gap of length equal to the current gap size count is incremented by one. The gap

size counter is then reset to zero, and the process continues.

Gap count data is accumulated in one hour blocks and then transferred to disk for permanent storage. By using this approach, in the event of a test failure, all but the last hour of data would be valid and available for use. In testing the status bits, the loss of synchronization signal is always tested first. LOS represents such a catastrophic error that the experiment is terminated when that signal is asserted.

The major limitation to this approach is the relatively slow processing speed of the microcomputer. When two or more errors occur in succession, the microcomputer may not be able to process the gap count data as fast as it is generated. For example, with a data rate of 1.544 Mbps, the time between the arrival of two successive error bits is about 650 nanoseconds. In that time period, the pc can perform only about four instructions. Even with interrupt driven software, this speed is inadequate to process the data.

For this reason, an interface between the BERT and the microcomputer is necessary. A block diagram of this Interface is shown in Figure 3. The interface consists primarily of a 32 bit counter and a 512 x 36 bit (only 32 bits are used) FIFO storage buffer. The counter is driven directly from the recovered clock from the BERT. Each time

a bit is received by the BERT, the counter is incremented by one. This counter represents the size of the current gap.

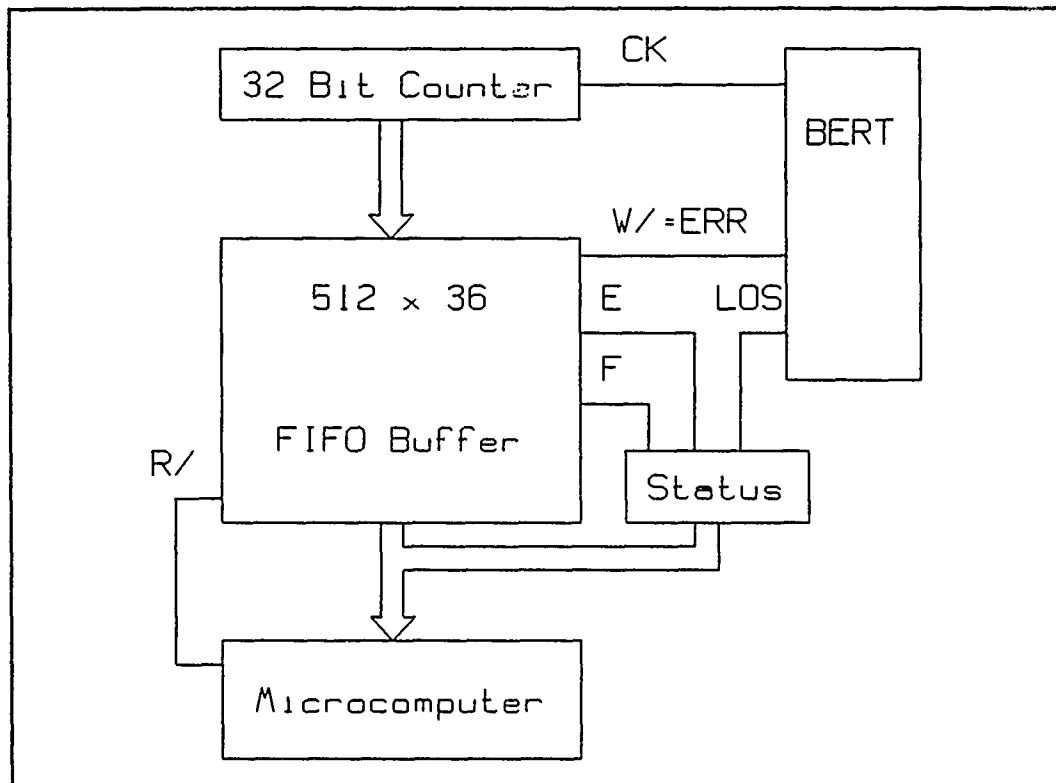


Figure 3: Computer Interface Block Diagram

When an ERR signal is received by the interface, this triggers two operations. First, the current value of the counter is written into the FIFO store for later retrieval by the microcomputer. Immediately thereafter, the value in the counter is cleared in preparation for beginning the next gap count sequence. All these actions are effected with a single inverter connected to the ERR signal, a silicon delay line, and a monostable multivibrator that generates the counter clear strobe.

The FIFO queue is a dual-ported RAM memory with integrated queue pointers. It contains internal circuitry

to provide a full (FF), empty (EF) and half-full (HF) flags. The CPU reads these flags as part of the data collection software. If EF is asserted, the CPU need do nothing and simply remains in the status read loop. If EF is negated, then the CPU must read data from the FIFO and move it into memory. Because the interface circuitry counts before writing, a count value of  $n$  in the register represents a gap of length  $n-1$ . The software makes an appropriate adjustment to the input data, either as it is read, or before updating the gap statistics.

The FIFO queue will permit collection of data as long as an error burst of length greater than 512 bits does not occur before the computer can read the data. Should this happen, the queue may become full, and data could be lost. This is extremely unlikely with any reasonable data error rates. Should the FIFO become full, it would be treated as a catastrophic failure of the channel, and the test discontinued.

The major difficulty in designing this interface circuit deals with the timing of the various operations within the unit. For example, the write to the FIFO must not take place while the counter is being incremented. Nor should the counter be cleared before the write operation (from the counter) to the FIFO queue has been completed. These criterion are shown in Figure 3. This figure includes the minimum timing requirements imposed by the various

components in the system, and also shows the timing values achieved by the interface circuit.

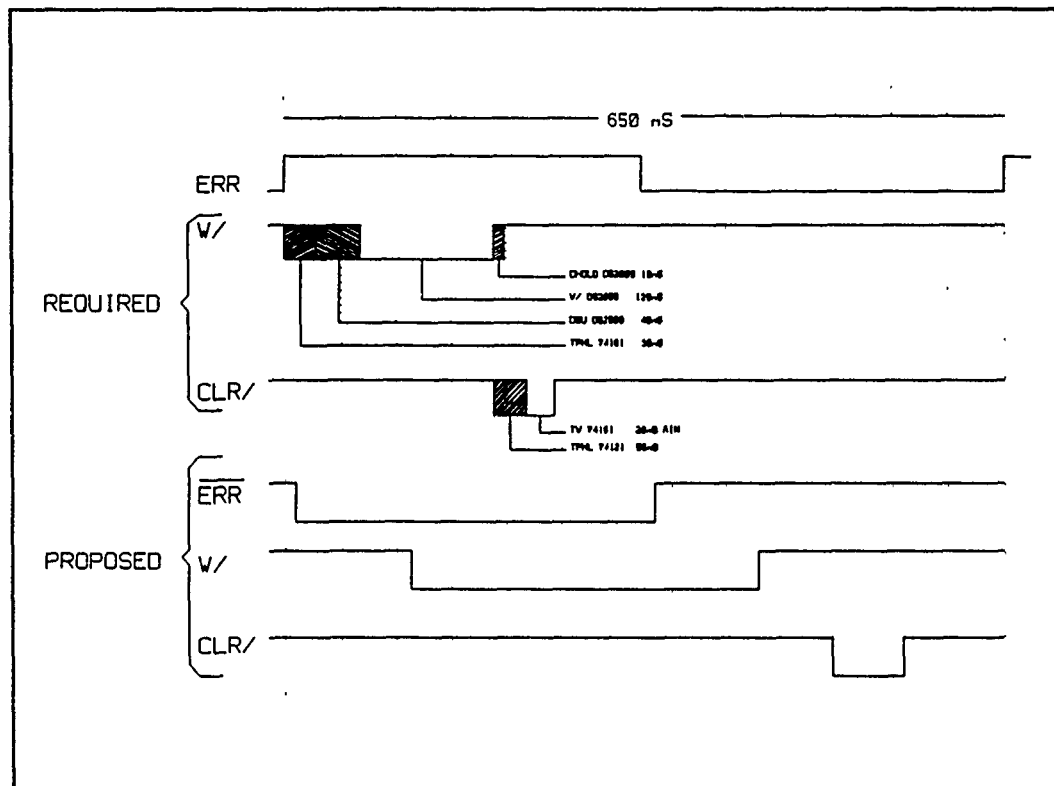


Figure 4: Interface Timing Requirements

The interface circuit is reasonably simple, and is constructed using wire wrapped circuitry on a JDR Microdevices brand prototyping card that includes address decoding. The inverter (74LS04) inverts the positive ERR signal to provide the active low W/ signal required to write to the FIFO. The silicon delay circuit (DS 1000-175) provides a 105 nanosecond delay to insure that the counter has stabilized before the write takes place (data set up time). The rising edge of the W/ signal triggers the monostable multivibrator (74LS122) to produce an

approximately 60 nanosecond low pulse to clear the 32 bit counter after the write has been completed.

At 1.544 Mbps data rates, the minimum arrival time between error pulses could be as short as 650 nanoseconds. All the operations noted above are completed within this time frame and before the next CK and/or ERR signals arrive. The average time between arrivals of error bits (i.e. the average gap length) is the reciprocal of the bit error rate (BER) of the channel being tested, and sequences of errors longer than 10 bits would be extremely rare. The FIFO store is capable of handling at least 512 errors in sequence without losing data. Should that prove to be inadequate, pin compatible FIFO stores with up to 8K storage locations are commercially available at a slightly increased cost.

The software for this experiment is also relatively straight forward. The microcomputer reads the status port to determine if the FIFO contains data. If so, the FIFO is read and the gap statistics are updated based on the value read. The computer then returns to the loop of reading the status port. If the status port indicates that the FIFO is empty, the computer continues to read the port until data is available.

Approximately once per hour, the CPU will write the accumulated error data to disk. To insure a complete record, this write operation is performed in conjunction

with a error occurrence in the data stream. This way, a single gap will not be recorded as two gaps of smaller size.

The technique of collecting and storing gap data rather than raw error data represents a compromise. The most comprehensive data would be obtained by storing the string of error bits, exactly as they occur over time. The problem with attempting to store the error information in this format is that the storage requirements would be enormous. The most practical method for storing the error data is to collect and store only the error gap data. This approach permits a relative simple program for keeping track of the gap counts, and also requires minimal storage for the data. Since not all gap sizes will occur in any particular experiment, a linked list is the most appropriate data structure for storing the gap data. The software for this system is currently under development. A block diagram for the full interface circuit is given in Figure 4.

#### IV. CONCLUSIONS:

The only real long term solution to the problem of providing Markov models for the EIUs is to collect the data and then convert the data to Markov models. The experiment designed for this purpose during this project represents an excellent beginning on this task. The hardware is currently complete, and the software under development should be capable of providing enough data to meet the needs of the



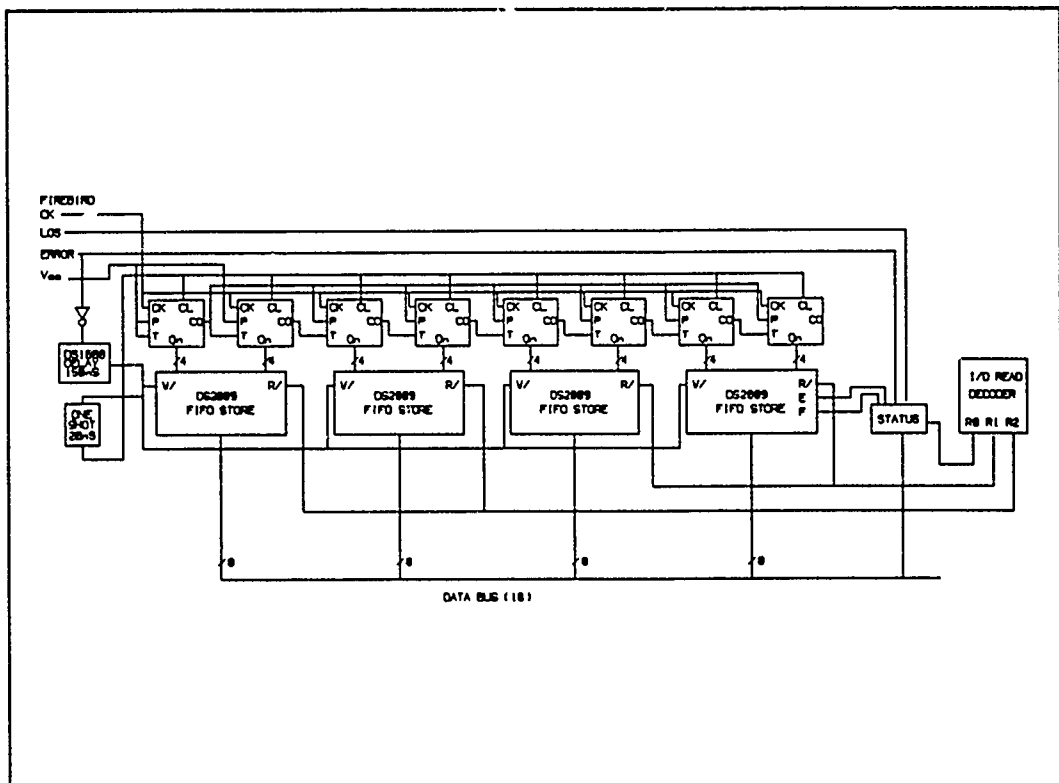


Figure 4: Interface Circuit

robust protocol project. This project should be continued, and data should be collected on a variety of data communications links under many different circumstances.

The current interface should probably be duplicated and/or improved as needed to permit the simultaneous collection of data on full duplex communications links. Additional units may be needed to permit simultaneous data collection on various media such as HF, VHF, UHF, etc. One desirable future goal would be to establish a HF data link between Mississippi State University and Rome Laboratory for the continued collection of data and for the extension of interaction between these two organizations.

Research into the use of heuristic models should continue. Currently, these models require more states than are available in the EIU. One possible future experiment would be to build computer simulation models of the heuristic models and use these programs to generate gap data. This gap data could then be converted to Markov models using the currently available models. The practicality and value of such models is worthy of further study. Such models may provide some insight into the nature of the error processes.

Some additional research should be devoted to a study of the potential for developing certain analytical models of data communications channels to determine if such models could be adapted for use in the EIU. Examples might included meteor burst and troposcatter channels.

#### V. REFERENCES:

- [1] Rome Research Corporation, Error Injector Unit (EIU) Critical Design Document, Facility Documentation, RADCRBC-90-031, 6 Mar 1990.
- [2] Tsai, S., "Markov Characterization of the HF Channel," IEEE Transactions on Communications Technology, Vol Com-17, pp24-32, Feb. 1969.
- [3] Gilbert, E. N., "Capacity of a Burst-Noise Channel," Bell Systems Technical Journal, Vol 39, pp 1253-1266, Sept. 1960.

- [4] Fritzman, B. D., "A Binary Channel Characterization Using Partitioned Markov Chains," IEEE Transactions on Information Theory, Vol. IT-13, pp 221-227, April 1967.
- [5] Fritzman, B. D. and Loenard, J. F., "Test Results of a Time-Dispersed Forward Error Control System," IEEE Transactions on Communications Technology, Vol. Com-13, pp 233-234, June 1965.
- [6] McManamon, Peter, "HF Markov Chain Models and Measured Error Averages," IEEE Transactions on Communication Technology, Vol Com-18, 00 201-208, June 1970.
- [7] Tsai, S., "Markov Characterization of the HF Channel", IEEE Transactions on Communication Technology, Vol Com-17, pp 24-32, Feb 69.
- [8] Tsai, S., "Simple Partitioned Markov Chain Model and Troposcatter Channel," Proceedings of the National Telecommunications Conference, 1973.
- [9] Varshney, Pramod, "Channel Models for The Error Injector Unit", Rome Air Development Center Technical Report RADC-TR-90-89, May 1990.
- [10] Kanal, L.N. and Sastry, A.R.K., Models for Channels with Memory and Their Applications to Error Control," Proceedings of the IEEE, Vol. 66, No. 7, July 78, pp 724-744.
- [11] Swarts, F., Ferreira, H.C. and Oostehuizen, "Renewal Models for PSK on Slowly Fading Rayleigh Channel,"

Electronic Letters, v 25, n 22, Oct 26, 1989, p 1514-1515.

- [12] Bargallo, J.M., "simulation of Convolutionally Encoded BPSK in Channels with Memory Modeled by MARKOV Chains, Unpublished paper, University of Kansas, 1991.
- [13] Vogler, Lewis E., "An Extended Model for Bit Error Statistics," NTIA Report 86-195, U.S. Department of Commerce, National Telecommunications and Information Agency, July, 1986.
- [14] Adoul, J-P A., Friotzman, B.D. and Kanal, L.N., "A Critical Statistic for Channels with Memory," IEEE Transactions on Information Theory, Vol IT-18, pp 133-141, Jan 72.
- [15] Turin, W., "Simulation of Error Sources in Digital Channels," IEEE Journal on Selected Areas in Communications, Vol 6, No 1, pp 85-93, Jan 88.
- [16] Ephremides, A. and Snyder, R.O., "Modeling of High Error Rate Binary Communications Channels," IEEE Transactions on Information Theory, Vol IT-28, pp 549-555, May 82.
- [17] Bussgane, J.J, Getchell, B.G. and Mahoney, P.F., "Stored Channel Simulation of Tactical VHF Radio Links," IEEE Transactions on Communications, Vol COM-24, No 2, pp 154-163, Feb 76.

# APPROXIMATING NEURAL NETS WITH $C^1$ NEURAL NETS

Michael D. Taylor, Associate Professor of Mathematics

## Abstract

Given a neural net whose architecture is defined by continuous functions, a method is exhibited for constructing a second neural net whose behavior approximates that of the first one arbitrarily closely and whose architecture is defined by continuously differentiable functions. This provides a means of "training" the first network by error back-propagation even in instances where back-propagation is not directly applicable. This in turn gives a tool for studying neural nets with "nonstandard" architectures.

## 1. Introduction

Neural nets have emerged as a possible alternative or complement to standard computer technology. They are believed to have a potential for performing such tasks as image processing and pattern recognition in a swifter and more satisfactory fashion than standard computers. One very important characteristic of neural nets and one reason for hoping they can perform some otherwise difficult tasks is that neural nets have the potential to be trained rather than having to be programmed.

The most widely used method of training neural nets is the error back-propagation algorithm. There is a certain technical restriction on this method: The algorithm requires that the ways in which units of a neural net depend on one another all be describable in terms of what are called continuously differentiable functions (also known as  $C^1$  functions). Not all nets meet this requirement, so back-propagation cannot be applied to them.

A particular motivation for this research was interest in neural nets having an architecture suggested by concepts occurring in fuzzy analysis. Fuzzy analysis is a method of

dealing with uncertainty, particularly uncertainty arising from vagueness rather than randomness. (Example: When is a pear ripe? There is more than one possible answer.) There has recently been a great deal of interest in using fuzzy logic in control problems; for example, as described in [3], an automatic train operating system has been successfully tested in Japan. Fuzzy analysis has also been turned to the problem of pattern recognition (see [2]) with possible applications to military intelligence and medical data. It should also be noted that there is recognition in the neural network community of the desirability of studying neural networks with nonstandard architectures. This point is specifically brought out in the DARPA study on neural networks, [1]. Another example occurs in [5] where nonstandard architectures of interest in image processing are mentioned as worthy of further investigation; this last example makes use of operations typical of fuzzy analysis.

Very little seems known about neural net architectures which derive from fuzzy analysis concepts. (But an example of such an architecture can be seen in [4].) One difficulty in investigating such architectures is that the functions one typically uses in fuzzy analysis (maximum and minimum) are nondifferentiable. This means the back-propagation algorithm, our most widely used tool in training a neural net, would not be applicable.

The objective of this work has therefore been to find a method — comparable to or analogous to the back-propagation algorithm — of training “fuzzy” neural nets. The outcome has been better than that. A way has been found to “extend” back-propagation to any neural network whose workings can be completely described in terms of continuous functions. This includes not only the “fuzzy” nets but presumably a large number of other types as well whose architectures have not even as yet been described.

## 2. The Problem

The back-propagation algorithm has been described many times in the literature on neural nets. (See, for example, [6] or [7].) This description has usually been given with certain

restrictions in mind, namely that computations of cell states will involve multiplying connection weights by other cell states, adding a number of results of this kind together, then adding (or subtracting) a "threshold" value, and finally applying a so-called "squashing function" to the result. These particular details do not matter. What does matter is that a very particular architecture is being assumed. In this work we still restrict ourselves to layered, feedforward networks, but for our purposes we need -- and give -- a description of the back-propagation algorithm which applies to more general architectures than those usually encountered in the literature.

We then come to the basic idea of this work. Suppose we are given a neural net whose workings are described by continuous functions. We construct a second neural net whose behavior approximates that of the first net very closely -- as closely as we may wish -- but whose workings are described by  $C^1$  (that is, continuously differentiable) functions. Let us call this second network a  $C^1$  network. The fact that the approximating  $C^1$  network can be constructed is a straightforward application of some well-known facts from mathematical analysis. A proof of the construction and a particular way to carry it out are given. The back-propagation algorithm can now be applied to train the  $C^1$  network. The results of this training -- connection weights and threshold values of cells -- can then be transferred to the original network. Since the two networks behave in very similar ways, the original network then behaves as though trained.

We then exhibit certain technical details which are useful in actually carrying out this program and conclude with some analysis of certain architectures to which back-propagation would not normally be applicable.

### 3. Error Back-propagation

It is convenient for our purposes to first describe a generalized form of error back-propagation for a synchronous, feed-forward, layered neural net. The discussion given below actually applies to slightly more general networks. We assume we have units (or neurons or cells)

$u_1, u_2, \dots, u_N$ . Let

$I$  = the set of  $i$  such that  $u_i$  is an input unit,

$O$  = the set of  $i$  such that  $u_i$  is an output unit,

$\mathcal{P}_i$  = the set of  $j$  such that there is a connection running from  $u_j$  to  $u_i$ ,

$s_i$  = the state of the  $i$ th unit,  $u_i$ ,

$w_{ij}$  = the weight of the connection running from the  $j$ th to the  $i$ th unit,

$\theta_i$  = the "bias" or "threshold" value of the  $i$ th unit.

We think of  $\mathcal{P}_i$  as the set of "predecessors" or "parents" of the  $i$ th unit, and we put bias and threshold in quotes in the definition of  $\theta_i$  because we may use it in ways which make it inappropriate to picture it as playing that role.

It is usual to assume the states of the units are related by an equation of the form

$$s_i = f_i \left( \sum_{j \in \mathcal{P}_i} w_{ij} s_j + \theta_i \right),$$

where the  $s_j$ 's are understood to be the states of the  $u_j$ 's at some time  $t$ , where  $s_i$  is understood to be the state of  $u_i$  at time  $t+1$ , and where  $f_i$  is a nonlinear function often referred to as a "squashing" function. However we shall assume the more general relation

$$s_i = s_i \left( \{s_j\}_{j \in \mathcal{P}_i}, \{w_{ij}\}_{j \in \mathcal{P}_i}, \theta_i \right)$$

where as before the  $s_j$ 's are understood to be states at time  $t$  and  $s_i$  is a state at time  $t+1$ . In this last equation  $\{s_j\}_{j \in \mathcal{P}_i}$  and  $\{w_{ij}\}_{j \in \mathcal{P}_i}$  are to be thought of as vectors whose components are indexed by  $\mathcal{P}_i$ . The  $s_j$ 's and  $w_{ij}$ 's and  $\theta_i$  play the role of independent variables in the function  $s_i$ , and it is very important to note that  $s_i$  is assumed to be a continuously differentiable function of these variables.

Let  $E$  be an index of performance used in training the neural net. We assume that

(1)  $E$  is to be minimized,

(2)  $E$  is a continuously differentiable function of the output states,  $E = E \left( \{s_i\}_{i \in O} \right)$ .



$E$  might be the sum of squares of training errors or some sort of entropy or some other sort of index. Error back-propagation gives us a method of incrementally changing the  $w_{ij}$ 's and  $\theta_i$ 's so as to produce corresponding, incremental decreases in  $E$  which will eventually (hopefully) force  $E$  to a minimum value or at least close to such a minimum.

We first take a fixed input vector  $\{s_i\}_{i \in I}$ . Since the output states  $\{s_k\}_{k \in O}$  are now completely determined by the weights and biases of the neural net, we may view  $E$  as a function of the weights and biases,

$$E = E(\{w_{ij}\}_{i,j}, \{\theta_i\}_i)$$

for this fixed input. Then the following equation essentially describes the first Taylor polynomial for  $E$ :

$$\delta E = \sum_{i,j} \frac{\partial E}{\partial w_{ij}} \delta w_{ij} + \sum_i \frac{\partial E}{\partial \theta_i} \delta \theta_i.$$

We wish to change the  $w_{ij}$ 's and  $\theta_i$ 's in such a way as to carry  $E$  in the direction of steepest decrease. Accordingly we set

$$\delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \quad \text{and} \quad \delta \theta_i = -\eta \frac{\partial E}{\partial \theta_i}$$

where  $\eta$  is a small, positive number. It follows that we must be able to find some way to calculate these partials of  $E$  in terms of the fixed input vector.

Now choose  $i$  and  $j$ . It follows from (2) that we must have

$$\frac{\partial E}{\partial w_{ij}} = \sum_{k \in O} \frac{\partial E}{\partial s_k} \frac{\partial s_k}{\partial w_{ij}}$$

and

$$\frac{\partial E}{\partial \theta_i} = \sum_{k \in O} \frac{\partial E}{\partial s_k} \frac{\partial s_k}{\partial \theta_i}.$$

Note that the partials of  $E$  with respect  $s_k$  (where  $u_k$  is an output unit) are all computable since  $E$  as a function of the  $s_k$ 's is given. Next it follows from

$$s_p = s_p(\{s_q\}_{q \in \mathcal{Q}_p}, \{w_{pq}\}_{q \in \mathcal{Q}_p}, \theta_p)$$

that for every  $u_p$  which is not an input unit, we have

$$\frac{\partial s_p}{\partial w_{ij}} = \frac{\partial s_p}{\partial w_{pj}} \text{ if } p = i,$$

$$\sum_{q \in \mathcal{Q}_p} \frac{\partial s_p}{\partial s_q} \frac{\partial s_q}{\partial w_{ij}} \text{ if } p \neq i,$$

and

$$\frac{\partial s_p}{\partial \theta_i} = \frac{\partial s_p}{\partial \theta_p} \text{ if } p = i,$$

$$\sum_{q \in \mathcal{Q}_p} \frac{\partial s_p}{\partial s_q} \frac{\partial s_q}{\partial \theta_i} \text{ if } p \neq i.$$

Of course if  $s_p$  is an input state, these partials are zero.

It follows from the last two equations that beginning with the input states one may recursively calculate every

$$\frac{\partial s_p}{\partial w_{ij}} \text{ and } \frac{\partial s_p}{\partial \theta_i}$$

and hence

$$\frac{\partial E}{\partial w_{ij}} \text{ and } \frac{\partial E}{\partial \theta_i}.$$

#### 4. Some Approximation Results

We consider layered, feed-forward neural nets. We wish to approximate the functions which transform the contents of the nodes on one layer of the network to a new set of contents on the next layer by  $C^1$  transformations. This defines a new net, one to which it is possible to apply the error back-propagation algorithm. The important thing is to show that it is possible to construct the new net in such a way as to be sure its behavior approximates that of the old net in some satisfactory fashion. The epsilon-delta conditions given below are typically associated with uniformly continuous functions or continuous functions defined over compact sets.

We state the propositions below without proofs.

Our first result shows we can approximate the transformation from one layer of a network to a single unit in the next layer. Note that if  $u = (u_1, u_2, \dots, u_N)$ , a point in  $\mathbb{R}^N$ , then

$$|u| = \sqrt{u_1^2 + u_2^2 + \dots + u_N^2},$$

the standard Euclidean norm. We will also find it convenient to employ the norm given by

$$\|u\| = \max_i |u_i|.$$

PROPOSITION 1. Let  $F: \mathbb{R}^D \rightarrow \mathbb{R}$  be a locally integrable function such that  $|F(x)| \leq M$  for all  $x$  in the domain of  $F$ . Suppose  $\epsilon$  and  $\delta$  are positive numbers such that  $|F(x_1) - F(x_2)| \leq \epsilon$  whenever  $\|x_1 - x_2\| \leq \delta$ . Let  $\phi$  be a real-valued function such that

$\phi$  is  $C^1$ ,

$\phi \geq 0$ ,

$$\int_{\mathbb{R}^D} \phi = 1,$$

and

$$\|u - v\| \leq \delta \text{ for all } u, v \in \text{supp } \phi$$

where  $\text{supp } \phi$  stands for the support of  $\phi$ . Let  $G = F * \phi$ . Then  $G$  satisfies the following:

- (1)  $G$  is  $C^1$ .
- (2)  $|G(x)| \leq M$  for all  $x$ .
- (3)  $|G(x) - F(x)| \leq \epsilon$  for all  $x$ .
- (4)  $|G(x_1) - G(x_2)| \leq \epsilon$  whenever  $\|x_1 - x_2\| \leq \delta$ .

NOTE 1: The boundedness of  $F$  is used here only to deduce the boundedness of  $G$  by the same bound. If the statements about boundedness were omitted, the proposition would still be true. There are two justifications for including such a condition. On the one hand it is physically reasonable; in a real world neural net, the values assigned to a unit cannot become infinitely positive or negative. On the other hand, if we consider neural nets with an architecture inspired by fuzzy analysis, we may want our units to take on values between 0 and 1, and it may

be important to know that if a net described by  $F$  has this property, then the "approximating" net described by  $G$  also has this property.

Our next proposition shows how to extend the approximation process to the transformation which takes one layer of the network to the next layer.  $D$  may be thought of as the number of units and connections and biases in one layer and  $R$  as the number of units in the succeeding layer.

PROPOSITION 2. Suppose  $F: \mathbb{R}^D \rightarrow \mathbb{R}^R$ . We may write  $F$  in the form  $(F_1, F_2, \dots, F_R)$ . Assume that each  $F_i$  is locally integrable and that for all  $i$  and all  $x$  we have  $|F_i(x)| \leq M$ . Suppose  $\epsilon$  and  $\delta$  are positive numbers such that  $\|F(x_1) - F(x_2)\| \leq \epsilon$  whenever  $\|x_1 - x_2\| \leq \delta$ . Let  $\phi_1, \phi_2, \dots, \phi_R: \mathbb{R}^D \rightarrow \mathbb{R}$  satisfy

each  $\phi_i$  is  $C^1$ ,

each  $\phi_i \geq 0$ ,

$\int_{\mathbb{R}^D} \phi_i = 1$  for each  $i$ , and

for all  $i$  we have  $\|u - v\| \leq \delta$  for all  $u, v \in \text{supp } \phi_i$ .

For each  $i$ , set  $G_i = F_i * \phi_i$  and let  $G = (G_1, G_2, \dots, G_R): \mathbb{R}^D \rightarrow \mathbb{R}^R$ . We then have the following conclusions:

(5)  $G$  is  $C^1$ .

(6)  $|G_i(x)| \leq M$  for all  $i$  and  $x$ .

(7)  $\|G(x) - F(x)\| \leq \epsilon$  and  $|G(x) - F(x)| \leq \sqrt{R} \epsilon$  for all  $x$ .

(8)  $\|G(x_1) - G(x_2)\| \leq \epsilon$  and  $|G(x_1) - G(x_2)| \leq \sqrt{R} \epsilon$  whenever  $\|x_1 - x_2\| \leq \delta$ .

NOTE 2: If  $H$  is a function of  $N$  variables, we say the  $i$ th variable is active provided we can find  $u$  and  $v$ , differing only in the  $i$ th variable, such that  $H(u) \neq H(v)$ . In constructing the convolutions  $F_i * \phi_i$  note that  $\phi_i$  may be taken to be a function of only the active variables of  $F_i$  and that the evaluation of  $(F_i * \phi_i)(x)$  may possibly be obtained by integrating over a lower dimensional space than  $\mathbb{R}^D$ . This can be a useful fact if the number of nodes in the neural net is

large because  $D$  is an upper bound for the number of nodes in a certain layer. If  $F_i$  has no active variables, then instead of manufacturing  $\phi_i$ , we can simply set  $G_i = F_i$ . Then (5) follows because  $F_i$  is a constant function, and (6), (7), and (8) are true trivially.

The following result should be thought of as applying to neural nets with  $L$  layers. Each  $F_i$  can be considered to be the function which takes the node values, the connection weights, and the biases associated with one layer and transforms them into the node values associated with the next layer.

PROPOSITION 3. Let  $F_1, F_2, \dots, F_L$  map thus:

$$\mathbb{R}^{D_0} \xrightarrow{F_1} \mathbb{R}^{D_1} \xrightarrow{F_2} \mathbb{R}^{D_2} \xrightarrow{F_3} \dots \xrightarrow{F_L} \mathbb{R}^{D_L}.$$

Assume each component of each  $F_i$  is locally integrable and that there is a given sequence of positive numbers  $\delta_0, \delta_1, \dots, \delta_L$  satisfying

$$\|F_i(x_1) - F_i(x_2)\| \leq \delta_i \text{ whenever } \|x_1 - x_2\| \leq \delta_{i-1}.$$

For each  $F_i$  manufacture a  $G_i$  in the manner described in Proposition 2 with  $\delta = \delta_{i-1}$  and  $\epsilon = \delta_i$ . Then each  $G_i$  is  $C^1$  and for all  $x \in \mathbb{R}^{D_0}$  we have

$$\|(G_L \circ G_{L-1} \circ \dots \circ G_1)(x) - (F_L \circ F_{L-1} \circ \dots \circ F_1)(x)\| \leq L \delta_L.$$

NOTE 3: One may give another version of this proposition in which the condition

$$\|F_i(x_1) - F_i(x_2)\| \leq \delta_i \text{ whenever } \|x_1 - x_2\| \leq \delta_{i-1}$$

is replaced by

$$|F_i(x_1) - F_i(x_2)| \leq \frac{\delta_i}{\sqrt{D_i}} \text{ whenever } |x_1 - x_2| \leq \delta_{i-1}$$

and the conclusion is

$$|(G_L \circ G_{L-1} \circ \dots \circ G_1)(x) - (F_L \circ F_{L-1} \circ \dots \circ F_1)(x)| \leq L \delta_L.$$

NOTE 4. One may, if one wishes, think of the last layer of the network as corresponding to a performance index  $E$ . These propositions then indicate how to construct a  $C^1$  network in

such a way that its performance index will approximate that of the original network as closely as one desires.

## 5. Construction and Training of $C^1$ Networks

Let us now consider some of the details involved in carrying out the program described in the last two sections.

Suppose that we have found a sequence  $\delta_0, \delta_1, \dots, \delta_L$  of the sort required by Proposition

3. Let  $K_i$  be the higher dimensional "cube"

$$K_i = [-\delta_k/2, \delta_k/2]^{A_i}$$

where  $A_i$  is the number of active variables of the state  $s_i$  of the  $i$ th unit  $u_i$  of the network. We choose  $\phi_i: \mathbb{R}^{A_i} \rightarrow \mathbb{R}$  such that

$$\phi_i \text{ is } C^1,$$

$$\phi_i \geq 0,$$

$$\int \phi_i = 1,$$

and  $\text{supp } \phi_i \subseteq K_i$ .

Notice that this last condition implies  $\|u - v\| \leq \delta_k$  for all  $u, v \in \text{supp } \phi_i$ .

We now introduce functions which define the values of nodes in the  $C^1$  network. Set

$$\bar{s}_i = s_i \text{ if } u_i \text{ is an input node}$$

$$s_i * \phi_i \text{ if } u_i \text{ is not an input node.}$$

(Note: If we take  $E$  to be our output, then we may want to compute a new performance index of the form  $\bar{E} - E * \phi$ . This  $\bar{E}$  may be taken as the performance index of the  $C^1$  network.)

Let  $u$  and  $v$  stand for vectors at which  $s_i$  can be evaluated, i. e., they are shorthand symbols for vectors of the form

$$(\{s_j\}_j \in \mathcal{Q}_i, \{w_{ij}\}_j \in \mathcal{Q}_i, \theta_i).$$

We now begin the training of the  $C^1$  network.

Step 1. Starting with the inputs  $\{\bar{s}_i\}_{i \in \mathcal{I}} = \{s_i\}_{i \in \mathcal{I}}$  and given connection weights  $\{w_{ij}\}$  and thresholds  $\{\theta_i\}$ , compute the outputs of the  $C^1$  neural net,  $\{\bar{s}_i\}_{i \in \mathcal{O}}$  and the vectors  $u$  at which each  $\bar{s}_i$  is evaluated.

Step 2. Check the performance index to see if the network is trained. Typically this may amount to seeing if  $E \leq \epsilon_0$  (or  $\bar{E} \leq \epsilon_0$ ) for some given  $\epsilon_0$ . (It may be necessary to check the performance index for a number of input vectors, not just for one.) If the  $C^1$  network is trained, then stop and transfer the connection weights  $w_{ij}$  and threshold values  $\theta_i$  from the  $C^1$  network to the original network. By Proposition 3 the outputs of the original network will now differ from those of the  $C^1$  network by at most  $\epsilon = L\delta_L$ . If the  $C^1$  network is not trained, then go to the next step.

Step 3. Use the computed outputs to evaluate  $\{\partial E / \partial \bar{s}_i\}_{i \in \mathcal{O}}$  (or  $\{\partial \bar{E} / \partial \bar{s}_i\}_{i \in \mathcal{O}}$ ).

Step 4. If  $u_i$  belongs to layer 0 (the inputs), take all partials of  $\bar{s}_i$  to be 0 for all  $i$ .

Step 5. Consider the highest layer  $k$  of the  $C^1$  network with the property that for all units  $u_j$  in a lower layer all partials of the form

$$\frac{\partial \bar{s}_j}{\partial \bar{s}_m}, \quad \frac{\partial \bar{s}_j}{\partial w_{pn}}, \quad \text{and} \quad \frac{\partial \bar{s}_j}{\partial \theta_p}$$

have been computed (where  $u_m$  lies in a lower layer than that containing  $u_j$  and  $u_p$  lies in a layer no lower than that containing  $u_j$ ). Then for all  $u_i$  in layer  $k$ , we compute

$$\left( \frac{\partial \bar{s}_i}{\partial \bar{s}_j} \right) (u) = \int_{K_i} s_i(v) \left( \frac{\partial \phi_i}{\partial s_j} \right) (u-v) dv \quad \text{if } j \in \mathcal{P}_i$$

$$0 \quad \text{if } u_j \text{ belongs to layer } k-1 \text{ and } j \notin \mathcal{P}_i$$

$$\sum_{m \in \mathcal{P}_i} \left( \frac{\partial \bar{s}_i}{\partial \bar{s}_m} \right) \left( \frac{\partial \bar{s}_m}{\partial \bar{s}_j} \right) \quad \text{if } u_j \text{ belongs to a layer lower than } k-1.$$

$$\left(\frac{\partial \bar{s}_i}{\partial w_{mj}}\right)(u) = \int_{K_i} s_i(v) \left(\frac{\partial \phi_i}{\partial w_{mj}}\right)(u-v) dv \text{ if } m=i \text{ and } j \in \mathcal{P}_i$$

$$0 \text{ if } m=i \text{ and } j \notin \mathcal{P}_i$$

$$\sum_{n \in \mathcal{P}_i} \left(\frac{\partial \bar{s}_i}{\partial \bar{s}_n}\right) \left(\frac{\partial \bar{s}_n}{\partial w_{mj}}\right) \text{ if } u_m \text{ belongs to a lower level than } k.$$

$$\left(\frac{\partial \bar{s}_i}{\partial \theta_j}\right)(u) = \int_{K_i} s_i(v) \left(\frac{\partial \phi_i}{\partial \theta_j}\right)(u-v) dv \text{ if } j=i$$

$$0 \text{ if } u_j \text{ belongs to layer } k \text{ and } j \neq i$$

$$\sum_{m \in \mathcal{P}_i} \left(\frac{\partial \bar{s}_i}{\partial \bar{s}_m}\right) \left(\frac{\partial \bar{s}_m}{\partial \theta_j}\right) \text{ if } u_j \text{ belongs to a layer below } k.$$

After having carried out these computations, if  $k < L$ , then repeat this step. If  $k = L$ , then go to the next step.

Step 6. For all  $i$  and  $j$  now compute

$$\frac{\partial E}{\partial w_{ij}} = \sum_{m \in \mathcal{O}} \left(\frac{\partial E}{\partial \bar{s}_m}\right) \left(\frac{\partial \bar{s}_m}{\partial w_{ij}}\right)$$

and

$$\frac{\partial E}{\partial \theta_i} = \sum_{m \in \mathcal{O}} \left(\frac{\partial E}{\partial \bar{s}_m}\right) \left(\frac{\partial \bar{s}_m}{\partial \theta_i}\right).$$

Step 7. For all  $i$  and  $j$  replace

$$w_{ij} \text{ by } w_{ij} - \eta \frac{\partial E}{\partial w_{ij}}$$

and

$$\theta_i \text{ by } \theta_i - \eta \frac{\partial E}{\partial \theta_i}$$



where  $\eta$  is some appropriately chosen small, positive number.

Step 2. Return to step 1.

## 6. Constructing Deltas from Epsilons

Propositions 1 and 2 refer to  $\epsilon$  and  $\delta$  which are related to one another in a certain way, and Proposition 3 refers to a sequence  $\delta_0, \delta_1, \dots, \delta_L$  with the property that each member of the sequence is related to the one before and after it in a certain way. We give here, without proof, a technical result which is sometimes helpful in constructing such pairs  $\epsilon$  and  $\delta$  and such sequences  $\delta_0, \delta_1, \dots, \delta_L$ .

First if  $F: K \rightarrow \mathbb{R}^n$ , where  $K \subseteq \mathbb{R}^m$ , recall that the  $i$ th variable of  $F$  is active if and only if there exist  $u$  and  $v$  in  $K$  which differ only in their  $i$ th components and have the property that  $F(u) \neq F(v)$ .

PROPOSITION 4. Let  $K$  be a compact, convex subset of  $\mathbb{R}^m$ . Suppose  $G_1, \dots, G_p: K \rightarrow \mathbb{R}^n$  are  $C^1$  maps and  $G: K \rightarrow \mathbb{R}^n$  is a continuous map with the property that for all  $x \in K$  there is an  $i$  such that  $G(x) = G_i(x)$ . Suppose further that each  $G_i$  is a function of at most  $A$  active variables. Then for all  $u, v \in K$  we have

$$\|G(u) - G(v)\| \leq A M \|u - v\|$$

where each  $G_i = (G_{i1}, G_{i2}, \dots, G_{in})$  and

$$M = \max \left\{ \left| \left( \frac{\partial G_{ik}}{\partial x_j} \right) (x) \right| : x \in K, 1 \leq i \leq p, 1 \leq k \leq n, \text{ and } 1 \leq j \leq m \right\}.$$

## 7. Some Case Analysis; Epsilons and Deltas

The principal difficulty in implementing the  $C^1$  approximation technique for neural nets lies in computing the epsilons and deltas, in particular the deltas of Proposition 3. Here we analyze how this could be done for a few particular cases.

Case 1. We consider a neural net with an architecture given by

$$s_i = f_T \left( \left\{ \sum_{j \in \mathcal{P}_i} w_{ij} s_j \right\} + \theta_i \right)$$

where  $f_T$  is the threshold function,

$$f_T(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } 1 < x. \end{cases}$$

Notice that  $f_T$ , though continuous, fails to be differentiable at 0 and 1 so that back-propagation is not directly applicable to this neural net.

Given  $\epsilon > 0$ , we wish to find  $\delta > 0$  such that  $\|u - v\| \leq \delta$  implies  $|s_i(u) - s_i(v)| \leq \epsilon$ .

We see that  $s_i$  is a continuous function which, depending on the values of its variables, is equal to one of the  $C^1$  functions

$$0, \quad \left( \sum_{j \in \mathcal{P}_i} w_{ij} s_j \right) + \theta_i, \quad \text{or} \quad 1.$$

Thus Proposition 4 is applicable.

Let  $B$  be a positive number which  $|s_m|$ ,  $|w_{mj}|$ , and  $|\theta_m|$  are "unlikely" to ever exceed for any possible  $m$  or  $j$ . It will be convenient to suppose  $B \geq 1$ . We restrict our attention to  $u$  and  $v$  satisfying  $\|u\|$  and  $\|v\| \leq B$ . We compute the  $M$  of Proposition 4:

$$\max_{i,j,m} \left\{ \left| \frac{\partial s_i}{\partial s_j} \right|, \left| \frac{\partial s_i}{\partial w_{mj}} \right|, \left| \frac{\partial s_i}{\partial \theta_j} \right| \right\} = \max_{i,j,m} \{ 0, |w_{ij}|, |s_j|, 1 \} = B.$$

Let  $A$  be the maximum number of active variables any  $s_i$  can have. Then by Proposition 4 we have

$$|s_i(u) - s_i(v)| \leq AB \|u - v\|.$$

Thus

$$\|u - v\| \leq \epsilon/AB \text{ implies } |s_i(u) - s_i(v)| \leq \epsilon.$$

Notice that this last inequality is independent of the choice of  $i$ .

Choose  $\epsilon > 0$  and let  $L$  be the number of layers of the network. Referring to Propositions 1 through 3, we see that we can set

$$\delta_L = \epsilon/L,$$

$$\delta_{L-1} = \delta_L/AB = \epsilon/LAB,$$

$$\delta_{L-2} = \delta_{L-1}/AB = \epsilon/LA^2B^2,$$

etc.

In general we have

$$\delta_k = \frac{\epsilon}{LA^L - k_B^{L-k}}.$$

These are the values needed to ensure that the outputs of the  $C^1$  network will differ from the corresponding outputs of the original network by at most  $\epsilon$ .

Case 2. We assume we have a network with architecture given by

$$s_i = \left( \bigvee_{j \in \mathcal{P}_i} (w_{ij} \wedge s_j) \right) \vee \theta_i$$

where  $\vee$  stands for "max" and  $\wedge$  stands for "min" and  $0 \leq s_m, w_{mn}, \theta_m \leq 1$  for all  $m$  and  $n$ .

This is the sort of architecture investigated in [4].

The last case shows us the pattern to be followed.

Note that for any particular choice of variables we must have  $s_i = w_{ij}$  or  $s_j$  or  $\theta_i$ . Thus the partials of  $s_i$  turn out to be 0 or 1. It follows that the  $M$  of Proposition 4 is 1. As usual, let  $A$  = the maximum number of active variables for any unit. Therefore

$$\|u - w\| \leq \epsilon/A \text{ implies } |s_i(u) - s_i(w)| \leq \epsilon.$$

Note that this last inequality is independent of  $i$ .

Let  $\epsilon > 0$  be given and  $L$  be the number of layers of the network. Consulting Propositions 1 through 3 again, we see that we want

$$\delta_L = \epsilon/L$$

$$\delta_{L-1} = \epsilon/LA$$

$$\delta_{L-2} = \epsilon/LA^2$$

etc.,

or, in general,

$$\delta_k = \frac{\epsilon}{LA^L - k}.$$

Case 3. Let us assume an architecture of the form

$$s_i = f\left(\left(\sum_{j \in \mathcal{P}_i} w_{ij}s_j\right) + \theta_i\right)$$

where

$$f(x) = \frac{1}{1 + e^{-\alpha x}}$$

for some  $\alpha > 0$ . Architectures of this sort are mentioned in [5] as being of interest in image processing.

It is straightforward to show — and turns out to be useful to know — that the maximum value of  $f'(x)$  is  $\alpha/4$ . Notice that for particular choices of the variables we must have  $s_i = f(w_{ij}s_j + \theta_i)$  for some  $j \in \mathcal{P}_i$ . The derivatives of these functions must have the form

$$f'(w_{ij}s_j + \theta_i) w_{ij},$$

$$f'(w_{ij}s_j + \theta_i) s_j,$$

$$f'(w_{ij}s_j + \theta_i),$$

or

$$0.$$

Supposing  $|s_j|$ ,  $|w_{ij}|$ , and  $|\theta_i| \leq B$  and  $B \geq 1$ , then the magnitudes of these derivatives will all be bounded by  $\alpha B/4$ . Thus we may take  $M = \alpha B/4$  when applying Proposition 4.

Assuming  $\epsilon > 0$  given, we then see we should take

$$\delta_L = \epsilon/L,$$

$$\delta_{L-1} = (4\epsilon)/(L\alpha AB),$$

$$\delta_{L-2} = (4^2\epsilon)/(L(\alpha AB)^2),$$

etc.,

(where  $A$  is of course an upper bound on the number of active variables per node) or, in general,

$$\delta_k = \frac{4^{L-k} \epsilon}{L(\alpha AB)^{L-k}}.$$

**Case 4.** Consider the architecture defined by

$$s_i = f_T \left( \left( \bigvee_{j \in \mathcal{Q}_i} (w_{ij} + s_j) \right) + \theta_i \right)$$

where  $f_T$  is the threshold function from Case 1. This is again an architecture of the sort described in [5] as being of interest in image processing.

For particular choices of the variables we must have

$$s_i = 0 \text{ or}$$

$$w_{ij} + s_j + \theta_i \text{ for some } j \text{ or}$$

$$1.$$

The derivatives of these functions are 0 or 1. The analysis for this case reduces to that for Case 2, and we obtain

$$\delta_k = \frac{\epsilon}{LA^{L-k}}.$$

### 8. Conclusions

We have shown how, in theory, one can approximate a neural net whose architecture is defined by continuous functions by a neural net whose architecture is defined by  $C^1$  functions. This leads to a technique for indirectly applying the error back-propagation algorithm to the first neural net. In addition a slightly generalized version of the back-propagation algorithm is exhibited. Therefore a large number of neural net architectures are opened for investigation using the back-propagation algorithm which would not normally be accessible.

Several further lines of investigation are suggested by these results:

(1) Software should be developed to test the applicability and efficacy of the techniques put forth here. Can networks be trained this way to cope with such standard tasks as the XOP problem or character recognition?

(2) Can these approximation techniques be extended to networks whose architectures are defined by discontinuous functions? For example, the signum function is used in a number of architectures. The norm used to measure approximations in this present work is essentially the  $l_\infty$  norm. One suspects that such an extension would require a different sort of norm.

(3) These techniques could be used to carry out an investigation of neural nets with "fuzzy" architectures (that is, architectures suggested by the concepts of fuzzy analysis). Again it would be reasonable to ask if such networks can be trained to handle such standard tasks as the XOR problem or character recognition.

(4) One of the difficulties in applying back-propagation is the occurrence of local minima which produce false solutions of the training problem. It is possible that the construction of  $C^1$  neural nets to approximate given neural nets might be used as a technique to get rid of such local minima.

(5) The version of the back-propagation algorithm presented here, combined with the idea of  $C^1$  approximations, might make it easier to investigate performance indices  $E$  other than the sum of the squares of the errors. For example, fuzzy analysis suggests the possibility of performance indices which are linked to the notion of inference or implication and which are constructed using nondifferentiable functions.

#### REFERENCES

1. DARPA Neural Network Study. Fairfax, VA: AFCEA International Press, 1988.
2. J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York and London: Plenum Press, 1981.

3. S. Miyamoto, S. Yasunobu, and H. Ihara, "Predictive fuzzy control and its application to automatic train operation systems," in J. C. Bezdek, ed., Analysis of Fuzzy Information, Vol. II. Artificial Intelligence and Decision Systems. Boca Raton, FL: CRC Press, Inc., 1987.
4. W. Pedrycz, "Neurocomputations in relational systems," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13 (1991), 289-297.
5. G. X. Ritter, D. Li, and J. N. Wilson, "Image algebra and its relationship to neural networks," SPIE, Vol. 1098, Aerospace Pattern Recognition, (1989), 90-101.
6. D. Rumelhart, D. Hinton, and G. Williams, "Learning internal representations by error propagation," in D. Rumelhart and F. McClelland, eds., Parallel Distributed Processing, Vol 1. Cambridge, MA: M.I.T. Press, 1986.
7. P. J. Werbos, "Backpropagation through time: What it does and how to do it," Proceedings of the IEEE, Vol 78 (1990), 1550-1560.

# **OPTICAL FIBER AMPLIFIERS AND OSCILLATORS**

**BY**

Kenneth J. Teegarden  
Salahuddin Qazi

**SEE: SALAHUDDIN QAZI**



# SIMULATION MODEL INTEGRATION METHODOLOGY FOR ROME LABORATORIES

JEFFREY D. TEW

*Virginia Polytechnic Institute and State University, Blacksburg, Virginia*

This document provides an outline for implementing a successful hierarchical, integrated simulation model at Rome Laboratories for the purpose of modeling tactical battle engagement scenarios. It is recommended that an analytical framework be used to guide the development of the integration process. This analytical framework focuses on: (a) the transfer of data elements between model modules, (b) the estimation of response population characteristics of interest, (c) the design of simulation experiments in order to estimate a regression model of interest, and (d) the application of variance reduction techniques throughout the integrated model. The result of fully integrating the existing simulation models would be the ability to *accurately* and *validly* model tactical battle engagement situations with unprecedented, varying levels of detail that could be selected according to the modeler's agenda. Thus, the fully integrated model would serve as a valuable test bed for evaluating: (a) existing and proposed hardware systems, (b) existing and proposed simulation models, and (c) battle engagement strategies.

*Key Words:* simulation, hierarchical model integration, control variates, linear models, simulation experiments, variance reduction techniques.

## 1. Introduction

In this report we briefly outline a methodology for fully integrating, in a hierarchical fashion, existing simulation models currently used at Rome Laboratories for modeling aspects of tactical battle engagement scenarios (although the basic elements of this methodology could be effectively used to integrate different levels of simulation models used in other modeling arenas that are of interest to the Air Force, such as communications network models, materiel and personnel traffic flow models, etc.). The result of fully integrating the existing simulation models would be the ability to *accurately* and *validly* model tactical battle engagement situations with unprecedented, varying levels of detail that could be selected according to the modeler's agenda. Thus, the fully integrated model would serve as a valuable test bed for evaluating: (a) existing and proposed hardware systems, (b) existing and proposed simulation models, and (c) battle engagement strategies.

The tactical battle engagement arena is recommended as a test bed for formulating and implementing the simulation model integration methodology for the following reasons: (a) simulation models within this arena (as opposed to communications network models, etc.) are of paramount interest to the Air Force, (b) highly sophisticated simulation models of many aspects of the tactical battle engagement environment are currently being used at Rome Laboratories (many of these simulation models were developed by government contractors with the direction and cooperation of Rome Laboratories personnel), and (c) initial ground work for implementing the integration methodology has already begun through the efforts of the Joint Simulation Panel (JPL) which was established in 1988 at Rome Laboratories primarily through the efforts of Joe Cruskie (COAA), Neal Marples (COAA), Jim Papagni (IRRA), and Alex Sisti Jr. (IRAE) and the recommendation of IR and CO. Currently, JPL consists of Jerry Dussault (COAA), Jim Papagni (IRRA), and Alex Sisti Jr. (IRAE).

Together, they have begun to formulate a means of directing and managing a truly integrated simulation effort at Rome Laboratories. Up until this time, their ideas for simulation

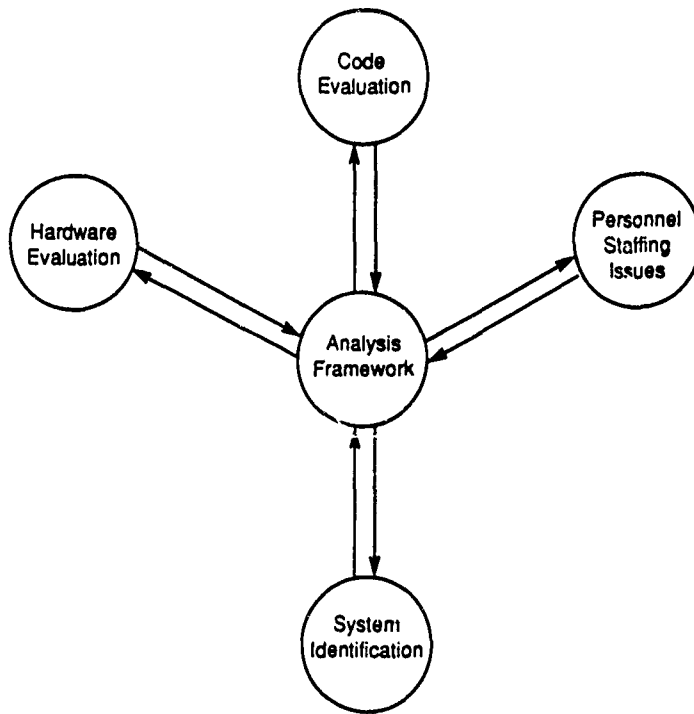
model integration have focused on: (a) personnel staffing and involvement issues regarding the simulation integration problem, (b) evaluation of the candidate simulation code so as to determine those simulation programs that will participate (at least initially) in the integration effort, (c) selection of a specific system scenario to be modeled initially, and (d) evaluation of the computer hardware that may serve as platforms for the integrated model.

However, through their efforts, they have identified one *key* component necessary to successfully developing an integrated simulation effort at Rome Laboratories that is currently missing. That component is an analytical (statistical) framework for directing the transfer of data (data blocks, statistical measures, etc.) between the model components that will, necessarily, comprise the integrated simulation model structure as well as coordinate the implementation of variance reduction techniques and performance measure estimation strategies that are critical to obtaining optimal information outputs (especially for large-scale simulation models). This analytical framework is the *key* component to the entire simulation integration effort because it will guide the efforts in the other four areas mentioned above to a common goal. This concept is indicated in Figure 1 below.

In the remainder of this document, we develop an outline for structuring the analytical framework for the simulation integration project. In the next section we outline the part of the analytical framework that will coordinate the transfer of data sets between simulation models within the integrated model. In Section 3 we discuss how the estimation of performance measure characteristics of interest should be structured within the analytical framework. and in Section 4 we indicate how the implementation of variance reduction techniques should be done within the integrated model. Together, these three sections offer a comprehensive structure for developing a successful integrated simulation model with unparalleled performance capabilities for analyzing force effectiveness.

## 2. Data Transfer

Consider the hierarchical simulation model configuration depicted in Figure 2. We anticipate



**Figure 1: Integrated Model Subtask Interaction**

four primary levels of model effort that, together, will comprise the fully integrated simulation model. These four levels are, from narrowest scope to broadest scope: (a) Primary Component Models (Level 1), (b) Elementary Systems Models (Level 2), (c) Small Mission Models (Level 3), and (d) Campaign Models (Level 4).

The Level 1 models primarily deal with modeling individual systems such as jammers, sensors, transmitters, etc.. They receive inputs directly from the model analyst, or some database indicated by the model analyst, and send outputs directly to the Level 2 models. They represent the highest level of detail in the integrated model. Currently, a number of simulation models at this level already exist at Rome Laboratories and are under frequent use.

Level 2 models focus on modeling the interaction of a Level 1 model with a specific platform; e.g., the modeling of an ESM suite installed on an aircraft. The effectiveness of the installed system is typically evaluated in the context of a one-on-one or few-on-few analysis. The inputs into models at this level are comprised of two types: (a) those received as outputs from the Level 1 models and (b) direct inputs from the model analyst. Models at this level retain all of the detail of the models at Level 1 via the Level 1 outputs that are fed directly into them. However, they incorporate this detailed information into a broader scope of engagement activities. Outputs from the Level 2 models are fed directly into Level 3 models. Currently, Rome Laboratories maintains an active development of models at this level (e.g., The Tac Brawler simulation package being used in modeling efforts in NCTI).

Level 3 models focus on modeling whole sub-theater engagements via the information fed into the model as outputs from the Level 2 models and the model analyst. Models at this level take an even broader view of an engagement scenario than Level 2 models, although still retaining, in theory, *all* of the detail contained at the lower levels. We anticipate that Level 3 models will be used to ascertain the effectiveness of tactical plans and decision making schemes in a combat mission environment. Currently, Rome Laboratories is beginning to use

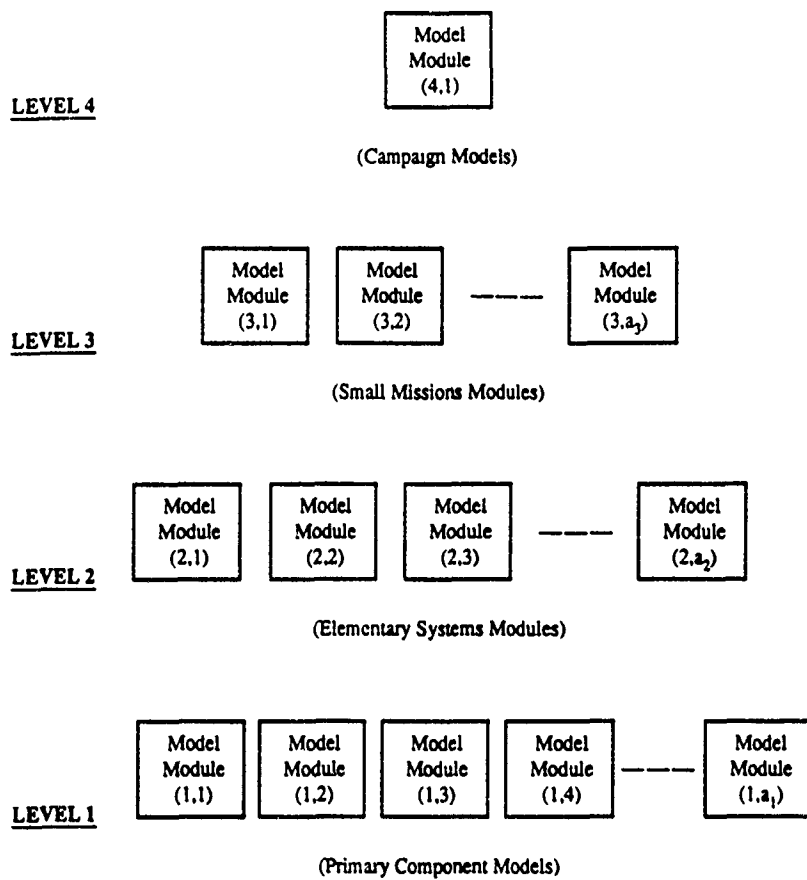


Figure 2: Hierarchical Simulation Model

a Level 3 model (Tac Suppressor) within NCTI. Outputs from the Level 3 models will be fed into the Level 4 model.

Finally, the Level 4 model represents the most comprehensive model of all. That is, it has the *broadest* scope in terms of its model domain. This model will encompass *all* relevant activity associated with operations in the joint (Air Force, Army, and Navy) campaigns against a full complement of combined enemy forces. The model will be used to test alternative strategic plans as well as high detail, lower level components within a full campaign scenario. Inputs to the Level 4 model will be obtained from the Level 3 models and the model analyst. Outputs from this model will go directly to the model analyst and  $C^3I$  for evaluation and decision making. Currently, Rome Laboratories does not possess this software capability. However, the analytical framework that we are proposing in this document will greatly serve in its development.

Next, we discuss, in general, how data is to be transferred between the models comprising the four levels. In order to do that, we first must define some useful notation. Let

- $x_{i,j,k}$  = the  $k$ th ( $k = 1, 2, \dots, b_{i,j}$ ) analyst supplied input to the  $j$ th ( $j = 1, 2, \dots, a_i$ ) model at the  $i$ th ( $i = 1, 2, 3, 4$ ) level.
- $w_{i,j,k}$  = the  $k$ th ( $k = 1, 2, \dots, c_{i,j}$ ) lower-level supplied input to the  $j$ th ( $j = 1, 2, \dots, a_i$ ) model at the  $i$ th ( $i = 2, 3, 4$ ) level.
- $y_{i,j,k}$  = the  $k$ th ( $k = 1, 2, \dots, d_{i,j}$ ) model generated output from the  $j$ th ( $j = 1, 2, \dots, a_i$ ) model at the  $i$ th ( $i = 1, 2, 3, 4$ ) level.
- $f_{i,j}$  = the conversion function that operates at the  $j$ th ( $j = 1, 2, \dots, e_i$ ) model of level  $i$  ( $i = 1, 2, 3, 4$ ).

The data generated by the model modules comprising the integrated simulation model will be of three types: (a) *time dependent information*, (b) *observational information*, and (c) *graphical information*. Time dependent information is dependent upon the length of the

simulation run. Observational information does not depend on the length of the simulation run. Graphical information will complement the first two forms of data, but will not add to it. That is, all relevant information will be given in one of the first two types listed above and the graphical data will be used to interface more effectively with the model analyst and other users.

The flow of information (data) utilizing these components is depicted in Figure 3. Solid lines with arrows indicate the direction of automatically fed data from one model level to another. Data is passed along these lines in standardized formats with prescribed agendas determined by the model from which the data originates. Dashed lines with arrows indicate how feedback information is routed back into a model at what is deemed, by the model analyst, an appropriate level. This feedback data process is monitored and administered by the model analyst by means of a decision process determined by the model analyst. Currently, we do not exactly know all of the information that will be made available to the model analyst during simulation runs and how often this information is updated (certainly, much of this information will be determined by the context in which the integrated model is to be built). We also do not know at this time all of the elements of the decision process used by the model analyst for interpreting this information. In fact, although it certainly would be expected that the actual analyst involvement would be accomplished via a computer program interface with the integrated model, at this point we offer no exact prescription for carrying out such a decision policy. However, we recommend that the *minimum* set of decisions to be made in the feedback phase by the model analyst consist of:

- Tactical simulation run decisions. These include determining when to start and to stop simulation runs for individual model modules. We anticipate that in a fully integrated simulation model that many of the individual model modules will start and stop at different times during the integrated model run. In large part, the starting and stopping of the individual model modules will be determined by conditions obtained



elsewhere in the integrated model. The analyst will be responsible for determining when these conditions are met for each model module in the integrated model and thus, implementing the implied action of the affected model module.

- Updating of individual model module estimates. This includes updating model estimates of relevance generated elsewhere in the integrated model and adjusting their input to the model module in question in a timely fashion. The analyst is responsible for coordinating the entire set of integrated model estimates made throughout the run of the simulation. This set includes estimates made by *all* model modules at *all* levels. That is, at any point during the simulation run the analyst is required to provide the most accurate and recent model estimates. This may also include the updating of any metamodel estimates of interest to the model analyst (see Section 3).
- Maintaining integrated model validity. The model analyst must maintain validity of the overall integrated simulation model. This is to be done on a continuing basis throughout the simulation run via periodically frequent updates on each model module.

For each of the three tasks listed above the analyst will have the ability to alter the frequency with which individual model modules are inspected so as to improve the run efficiency of the integrated model as well as the accuracy level of the model estimates. That is, if, for a particular module, the analyst determines that it only needs to "view" the module every  $2\delta t$  time units instead of, say, the original  $\delta t$  time units then that particular module efficiency will be improved. Similarly, if obtaining data from a particular model module every  $\delta t$  time units yields insufficient accuracy for a specific estimate, then the model analyst will be able to increase the frequency at which the model data is generated to, say,  $.5\delta t$  time units; thereby improving estimator accuracy. In other words, one may look upon the process of obtaining a data point from a lower-level model as taking a "snapshot" of that model and, at that point in time, retrieving the desired information. This process is repeated in

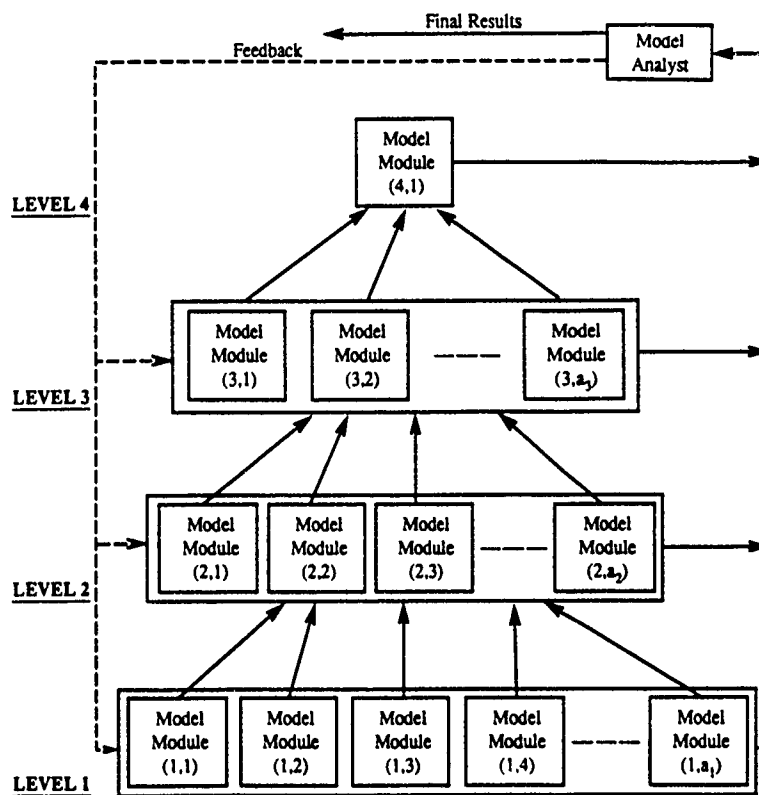


Figure 3: Data Flow

order to accumulate a database from that model. The quality of the statistical estimates constructed from that database depends, in large part, upon when and how frequently the data were collected. Consequently, giving the model analyst the power to adjust when the data are collected will allow him to adjust for the quality of the desired estimates. We anticipate, and recommend, that understanding and clearly identifying all of these roles that the model analyst will play in the integrated modeling process will be a critical part of the overall modeling effort. Indeed, the model analyst's role in the integrated model is the most important element for it will directly determine the quality of the entire simulation. Thus, it is imperative that this component of the integrated model be fully researched and understood. Currently, Rome Laboratories has devoted little effort to this portion of the integrated model effort. In fact, this aspect of the integrated model is new to the simulation community too and very little work has been done in this area. Consequently, we believe that it is very important to use the analysis framework so as to provide a sound and scientific means of guiding Rome Laboratories in piecing together the many parts of the integrated simulation model. Next, we discuss how the statistical estimation processes should be conducted.

### 3. Statistical Estimation

In this section we outline the statistical estimation procedures within the integrated, hierarchical simulation model framework. Identification and use of proper statistical estimates is critical to obtaining a clear understanding of the integrated model results.

There are two types of statistical estimation problems that will be encountered by the integrated simulation model. They are: (a) point and confidence interval estimation for a given output population  $y_{ijk}$  and (b) regression model (metamodel) estimation for a given response set of responses  $y_{jk}$  and their associated sets of inputs  $x_{ijk}$  and  $w_{ijk}$ . Both types of estimation problems need to be considered for *univariate* and *multivariate* cases since both may be encountered at any model module in the integrated model. Here, we consider only the univariate estimation type problems and recognize that many of the steps outlined herein

can be adapted to the multivariate case. Properly indentifying the necessary procedures for conducting statistical estimation in the simulation context will help to ensure proper interpretation of the results and more accurate measures of performance.

Unfortunately, statistical estimation (rarely an easy problem in itself) is made more difficult when applied to the simulation context. This is the result of the following conditions: (a) the sequence of outputs,  $y$ , obtained from a simulation run are often correlated and (b) the sequence of outputs,  $y$ , obtained from a simulation run are often *not* identically distributed. Under these two conditions statistical estimation becomes extremely difficult. However, there are some remedies that have been developed in the simulation literature that can help overcome some aspects of these difficulties. Unfortunately, they are not universally applicable and often do not provide a consistent measure of improvement across all types of simulation situations.

First, we consider the problem of estimating the mean and variance of the populations of the responses of interest. Running a computer simulation program can be viewed as generating a sample of observed responses values (for each response variable  $y_{ijk}$ ). Typically, the characteristics of interest for this population are the mean  $\mu_{ijk}$  and the variance  $\sigma_{ijk}^2$ . Because of the problems mentioned at the beginning of this section with regard to simulation data, the estimation of these two population parameters cannot be done well using classical statistical estimation techniques. For this reason, we provide the following outline for conducting statistical estimation for each  $y_{ijk}$  within the integrated model:

- For each response,  $y_{ijk}$ , of interest, determine the effect of initialization bias on the mean  $\mu_{ijk}$ . This would entail establishing a warm-up period for each model module in the integrated model and truncating the initial "biased" data obtained from this warm-up period. (Note that there may be different warm-up periods for different responses within a single model module.) Proper establishment of these warm-up periods will help to ensure that the estimates of the  $\mu_{ijk}$  will be unbiased. Currently, there are

several algorithms for determining the optimal initial warm-up period under different conditions.

- Determine how “batching” is to be done for each response  $y_{ijk}$  in the model during the simulation run. Identification of a proper batch size will reduce the effects that the dependency structure present in the sequence of observed responses and, thus, help to ensure unbiased estimation of the population variances  $\sigma_{ijk}^2$ . As with the initialization bias problem above, there may be different recommended batch sizes for *each* response,  $y_{ijk}$ , in the integrated model. There are several algorithms for establishing optimal batch sizes under various simulation contexts.
- Establish algorithms for determining optimal run lengths for each of the model modules comprising the integrated model. This would necessarily involve utilizing the estimation information obtained from both the estimation of  $\mu_{ijk}$  and  $\sigma_{ijk}^2$  for all response populations and performing a sequentially updated set of comparisons to the established accuracy levels indicated by the model analyst prior to simulation. Establishment of such a set of algorithms would greatly improve the efficiency of the entire integrated model.

In addition to these three primary steps to performing proper estimation during the simulation, rules should also be established for identifying how the multiple estimation problems within the integrated model context are interrelated.

Next, we provide an outline for conducting regression model estimation within the integrated model context. First, we introduce some basic notation to help clarify the issues.

Consider a simulation experiment consisting of  $m$  design points, where each design point is identified by the  $d$ -dimensional vector  $\varphi$  of design variables (factors) that are deterministic inputs to the simulation model. Let the univariate response from the  $g$ th design point be denoted by  $y_g$  and let the vector of responses from all  $m$  design points be denoted by

$y = (y_1, y_2, \dots, y_m)'$ . Also, let  $\varphi_g$  denote the settings of the  $d$  factors for the  $g$ th design point and let  $\{x_h() : h = 1, 2, \dots, p-1\}$  represent known functions of the factor settings. Then, assuming that the relationship between the response and the given functions of the factor settings is linear in the unknown parameters, we can write

$$y_g = \beta_0 + \sum_{h=1}^{p-1} \beta_h x_h(\varphi_g) + \epsilon_g \quad \text{for } g = 1, 2, \dots, m, \quad (1)$$

where  $\{\beta_h : h = 0, 1, \dots, p-1\}$  are the unknown model parameters and  $\epsilon_g$  represents the inability of the term  $(\beta_0 + \sum_{h=1}^{p-1} \beta_h x_h(\varphi_g))$  to determine  $y_g$ . Define  $\mathbf{X}$  to be the  $m \times p$  design matrix whose first column is all ones and whose  $(g, h+1)$  element is  $x_h(\varphi_g)$  ( $g = 1, 2, \dots, m$  and  $h = 1, 2, \dots, p-1$ ). Thus, the relationship between the response and the functions of the factor settings across all  $m$  design points can be written compactly as the following general linear model:

$$y = \mathbf{X}\beta + \epsilon, \quad (2)$$

where  $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})'$  and  $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_m)'$ .

If the  $\{x_h\}$  are chosen such that  $\mathbf{X}$  is orthogonal then:

$$\mathbf{X}'\mathbf{X} = m\mathbf{I}_p. \quad (3)$$

This can be achieved by a simple reparameterization, or coding, of the functions of the factor levels.

We also assume that

$$\{y_i : i = 1, 2, \dots, r\} \text{ IID } \sim N_m(\mathbf{X}\beta, \Sigma), \quad (4)$$

where  $\Sigma$  is the  $m \times m$  covariance matrix of  $y_i$  (for  $i = 1, 2, \dots, r$ ).

For our purposes, the design factors indicated in the above discussion would be comprised of both  $x_{ijk}$  and  $w_{ijk}$  inputs. Basically, these regression models would be used to approximate the relationships between a particular response variable of interest at a given

model module and the set of input factors that are fed into that model module. Knowledge of such a regression relationship can be used for any one of three powerful purposes: (a) control, (b) prediction, and (c) model validation. Thus, the regression relationship could be used to predict results without having to, necessarily, perform a simulation run. It could also be used to control certain modules so that reasonable outputs estimates could be obtained for them without having to run that particular module during the simulation run of the integrated model. This has the potential of fantastically reducing the needed CPU time for a particular run or set of runs. Lastly, the regression model could be used to help the model analyst validate portions of the integrated system. All of these uses for regression models fitted to simulation output have been well documented in the literature. However, to date, no one has attempted to simultaneously implement all such methods in one large scale simulation model such as the integrated model context here.

In the following, we list the recommended steps necessary for properly conducting a designed simulation experiment for the purpose of estimating a regression function of the type indicated in equation (1). Again, these steps would have to carried out for each set of inputs and the associated response of interest for all model modules.

- Determine which relationships are to be approximated by a regression model.
- Hypothesize a functional form for each model of interest.
- Determine the experimental design that is to be used.
- Validate the key assumptions made in equation (4) above.
- Estimate the unknown parameters in the model.
- Test for the adequacy of the fitted model.

Successful completion of these steps will provide the model analyst with a valid regression approximation that may be used for any of the three purposes above. Currently, it

remains unclear as to all of the uses such regression models may have. This aspect should be of great interest to the model analyst and has the potential for great payoffs.

In this section we have given an outline of the steps necessary in order to conduct good estimation routines within the model modules of the integrated simulation model. The primary goal of the model analyst should be to fully integrate these steps for each model module into the larger agenda of the entire integrated model. The level of coordination required to accomplish this task is considerable and has not been done previously. However, proper definition and understanding of all relevant terms as well as a clear view of the desired outcome can greatly facilitate in bringing it about.

#### 4. Variance Reduction Techniques

In this section we provide a menu of variance reduction techniques (VRT) that we think are essential to the integrated model effort. Variance reduction techniques are analysis methods that, under certain general conditions, can greatly reduce the variance of the estimators used in a simulation model without adversely affecting other aspects of the estimation issue (such as estimator bias). Typically, variance reduction techniques do not involve significant additional computational effort and, thus, can be thought of as gleaning more information from the simulation run without extra runtime costs.

Basically, variance reduction techniques include: (a) antithetic variates, (b) control variates, and (c) common random numbers. Each of these methods involves inducing correlations of a prescribed sign between the responses from a simulation model. The idea is to use this induced correlation to further improve the estimation of the population characteristic (mean or variance) of interest. Typically, researchers have focused attention on the implementation of these techniques to single simulation runs. However, recent attention has begun to focus on utilizing them across multiple run simulation experiments. This is of particular relevance to the integrated simulation model for two reasons. First, by nature, the integrated simulation model will involve multiple simulation runs from the various model



modules that comprise the integrated model. Second, many variance reduction applications have suggested that their effectiveness becomes enhanced for large-scale simulation models. Clearly, the integrated simulation model will be a large-scale simulation model and, hence, the potential will exist for reaping large benefits from the application of these techniques to the integrated model environment.

Consider the simple simulation experiment discussed in the previous section, where  $r$  represents a random sample from the uniform distribution on the unit interval  $[0, 1]$ . For a single replication of the basic  $m$ -point experimental design, we represent the complete set of, say  $k$  random number streams in the following way: (a) the sequence of random numbers available from the  $j$ th stream at the  $g$ th design point is

$$\mathbf{r}_{gj} = (r_{gj1}, r_{gj2}, \dots)'$$

( $g = 1, 2, \dots, m$  and  $j = 1, 2, \dots, k$ ); (b) the set of streams at the  $g$ th design point is

$$\mathbf{R}_g = (\mathbf{r}_{g1}, \mathbf{r}_{g2}, \dots, \mathbf{r}_{gn}) \quad (5)$$

( $g = 1, 2, \dots, m$ ); and (c) the aggregate random number input for the basic  $m$ -point experimental design is

$$\mathbf{R}^* = (\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_m).$$

Now at the  $g$ th design point,  $\mathbf{R}_g$  completely determines the events of the associated simulation run so that we can write

$$y_g(\mathbf{R}_g) = \beta_0 + \sum_{h=1}^{p-1} \beta_h x_{..}(\varphi_g) + \epsilon(\mathbf{R}_g). \quad (6)$$

In conducting a simulation experiment, the analyst must assign a particular set of random number streams  $\mathbf{R}_g$  to each design point  $g$  ( $g = 1, 2, \dots, m$ ). Three conventional assignment strategies are: (a) independent streams (the method of independent replications), (b) common streams (the method of common random numbers), and (c) antithetic streams (the method of antithetic variates). One strategy that assigns both antithetic and common

random number streams to the design points is described in the following. It has been shown to be optimal for a large class of simulation experiments.

Assume that the design matrix  $\mathbf{X}$  is *orthogonally blockable* into two blocks. That is, there exists an  $m \times 2$  matrix  $\mathbf{W}$  of zeros and ones such that: (a) design point  $g$  is in block  $j$  if and only if  $w_{gj} = 1$  ( $g = 1, 2, \dots, m$ , and  $j = 1, 2$ ); (b) the columns of  $\mathbf{W}$  are orthogonal to the columns of  $\mathbf{T}$  so that  $\mathbf{T}'\mathbf{W} = (\mathbf{0}_{p-1}, \mathbf{0}_{p-1})$ ; and (c) we have  $\mathbf{1}_m'\mathbf{W} = (m_1, m_2)$ , where  $m_1$  and  $m_2$  are positive integers respectively representing the size of each block so that  $m_1 + m_2 = m$ . For this situation we may use the following assignment rule:

If the  $m$ -point experimental design admits orthogonal blocking into two blocks of sizes  $m_1$  and  $m_2$ , preferably chosen to be as nearly equal in size as possible, then for all  $m_1$  design points in the first block, use a set of pseudorandom numbers  $\mathbf{R} = (r_1, r_2, \dots)$ , chosen randomly, and for all  $m_2$  design points in the second block, use the corresponding set of complementary pseudorandom numbers  $\bar{\mathbf{R}} = (1 - r_1, 1 - r_2, \dots)$ .

It has been shown that this strategy yields optimal results when the block sizes are equal; that is,  $m_1 = m_2$ . This strategy involves the decomposition of the error term  $\epsilon(\mathbf{R}_g)$  at the  $g$ th design point into a *random block effect*  $b(\mathbf{R}_g)$  and a residual  $\epsilon^\circ(\mathbf{R}_g)$ , both of which are functions of  $\mathbf{R}_g$ . Thus, the metamodel in (1) can be written as

$$y_g(\mathbf{R}_g) = \beta_0 + \sum_{h=1}^{p-1} \beta_h x_h(\varphi_g) + b(\mathbf{R}_g) + \epsilon^\circ(\mathbf{R}_g) \quad \text{for } g = 1, 2, \dots, m, \quad (7)$$

to emphasize the dependence of the random components in this metamodel on the random input  $\mathbf{R}_g$ . In addition, we make the following assumptions about the decomposition of the metamodel error terms in equation (7): (a) the errors  $\{\epsilon(\mathbf{R}_g)\}$  have mean zero and a constant variance  $\sigma^2$  across all design points; (b) the block effects  $\{b(\mathbf{R}_g)\}$  have mean zero and are not correlated with the residuals  $\{\epsilon^\circ(\mathbf{R}_g)\}$ ; (c) the residuals  $\{\epsilon^\circ(\mathbf{R}_g)\}$  have mean zero and are pairwise uncorrelated; (d) there is a nonnegative correlation between two block

effects obtained with common random numbers; and (e) there is a nonpositive correlation between two block effects obtained with antithetic (complementary) random number streams. These assumptions imply the following basic properties of the Schruben-Margolin correlation-induction strategy:

1. The response variance is constant across all points in the design so that

$$\sigma_g^2 = \text{var}[y(\mathbf{R}_g)] = \sigma^2 \quad \text{for } g = 1, 2, \dots, m. \quad (8)$$

2. There is a constant nonnegative correlation between all pairs of responses  $y_g$  and  $y_k$  ( $g \neq k$ ) that are realized from a common random number stream  $\mathbf{R}$ . Thus if  $\mathbf{R}_g = \mathbf{R}_k = \mathbf{R}$ , then

$$\text{corr}[y_g(\mathbf{R}), y_k(\mathbf{R})] = \rho_1 \quad \text{for } 1 < g, k < m \text{ and } g \neq k, \quad \text{where } 0 \leq \rho_1 \leq 1. \quad (9)$$

3. There is a constant nonpositive correlation between all pairs of responses  $y_g$  and  $y_k$  ( $g \neq k$ ) that are realized from antithetic random number streams  $\mathbf{R}$  and  $\bar{\mathbf{R}}$  respectively. Thus if  $\mathbf{R}_g = \mathbf{R}$  and  $\mathbf{R}_k = \bar{\mathbf{R}}$ , then

$$\text{corr}[y_g(\mathbf{R}), y_k(\bar{\mathbf{R}})] = \rho_2 \quad \text{for } 1 < g, k < m \text{ and } g \neq k, \quad \text{where } -1 \leq \rho_2 \leq 0. \quad (10)$$

Then, under these assumptions, with equal block sizes, the metamodel in (2) takes the following form:

$$\mathbf{y}(\mathbf{R}^{o*}) = \mathbf{X}\beta + \mathbf{W}\mathbf{B}(\mathbf{R}^{o*}) + \epsilon^o(\mathbf{R}^{o*}), \quad (11)$$

where:  $q \equiv \frac{m}{2} = m_1 = m_2$  is the common block size;  $\mathbf{B}(\mathbf{R}^{o*}) = [b_1(\mathbf{R}^{o*}), b_2(\mathbf{R}^{o*})]'$  is the  $2 \times 1$  vector of random block effects;  $\mathbf{W}$  is the  $m \times 2$  block incidence matrix; and  $\epsilon^o(\mathbf{R}^{o*})$  is the  $m \times 1$  vector of residual errors. Note that within each block, we assume a common block effect that *does not* depend on the design point. Let  $\mathbf{X}_j$  represent the  $\frac{m}{2} \times p$  design matrix for the  $j$ th block ( $j = 1, 2$ ). If the experimental points are so arranged that  $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2)'$ , then we have

$$\mathbf{W} = \begin{bmatrix} \mathbf{1}_q & \mathbf{0}_q \\ \mathbf{0}_q & \mathbf{1}_q \end{bmatrix} \quad (12)$$

where each column of  $\mathbf{W}$  contains  $q = \frac{m}{2}$  ones. With the assumptions given above, we have

$$\text{Cov}(\mathbf{B}) = \sigma^2 \begin{bmatrix} \rho_1 & \rho_2 \\ \rho_2 & \rho_1 \end{bmatrix}. \quad (13)$$

Expressions (12) and (13), together with the assumptions stated above, result in the following positive definite covariance structure for  $\mathbf{y}$ :

$$\Sigma_{\mathbf{y}}^{sm} = \sigma^2 \left[ \begin{array}{cccc|cccc} 1 & \rho_1 & \dots & \rho_1 & \rho_2 & \rho_2 & \dots & \rho_2 \\ \rho_1 & 1 & \dots & \rho_1 & \rho_2 & \rho_2 & \dots & \rho_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \dots & 1 & \rho_2 & \rho_2 & \dots & \rho_2 \\ \hline \rho_2 & \rho_2 & \dots & \rho_2 & 1 & \rho_1 & \dots & \rho_1 \\ \rho_2 & \rho_2 & \dots & \rho_2 & \rho_1 & 1 & \dots & \rho_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_2 & \rho_2 & \dots & \rho_2 & \rho_1 & \rho_1 & \dots & 1 \end{array} \right] = \sigma^2 \left[ \begin{array}{c|c} \Sigma_{11}^{sm} & \Sigma_{12}^{sm} \\ \hline \Sigma_{21}^{sm} & \Sigma_{22}^{sm} \end{array} \right], \quad (14)$$

where  $\det(\Sigma_{\mathbf{y}}^{sm}) \neq 0$ ,  $\Sigma_{11}^{sm}$ ,  $\Sigma_{12}^{sm}$ ,  $\Sigma_{21}^{sm}$ , and  $\Sigma_{22}^{sm}$  are  $q \times q$  matrices.

The following theorem pertains to simulation experiments performed under the strategy described above:

**Theorem 1:** If an experimental design admits orthogonal blocking, and if the assumptions of (8), (9), and (10) hold, then under the assignment rule the ordinary least-squares estimator of  $\beta$  has a smaller generalized variance than it has under the following strategies; (a) the assignment of one common set of random number streams to all design points, or (b) the assignment of independent sets of random number streams to each design point, provided

$$[1 + (m-1)\rho_1 - (2m_1m_2(m^{-1}))(\rho_1 - \rho_2)](1 - \rho_2)^p < 1,$$

in the latter case.

**Corollary 1:** Under the assumption of Theorem 1, the assignment rule is superior to the use of common random numbers in estimating  $\beta_0$ ; and the two strategies are equivalent in terms of dispersion for estimating  $\beta_1 = (\beta_1, \beta_2, \dots, \beta_{p-1})'$ .

When compared to the use of independent random number streams at each point, both the assignment rule and common random numbers are superior in terms of dispersion for estimating  $\beta_1$ . Thus, this strategy is an efficient means of combining the two correlation methods of common random numbers and antithetic variates for a large class of experimental designs.

The presentation given above indicates how some variance reduction techniques can be successfully applied across multiple simulation runs which constitute a simulation experiment.

In the context of the integrated model, the model analyst must perform many simulation experiments simultaneously (at least one multiple run experiment per model module). Consequently, the model analyst should focus much attention on how to successfully design and implement multiple run variance reduction techniques. To that end we give the following steps outlining what must be done.

- Identify all random components in each model module.
- Utilize random number generators that facilitate the synchronized use of all random numbers in every model module across the entire integrated model.
- Select, for each regression model that is to be estimated (see Section 3) and in conjunction with the design matrix, a random number assignment strategy so as to maximize model estimator performance.
- For each response of interest,  $y_{ijk}$ , select an optimal set of concomitant set of variables (control variates) that are highly correlated with the response. These "control variates" are then used to "explain away" or "control" some of the variability in the original response.
- Combine the use of the two techniques indicated by the previous two items.

Successfully implementing these steps will yield superior performance throughout the integrated model. Unfortunately, not all of these items have completely prescribed measures for being carried out. In addition, the model analyst's problems are further complicated when consideration is given to coordinating these various variance reduction steps from one model module to another. Nevertheless, enough is known about these variance reduction techniques to indicate that *unparalleled* benefits from fully integrating them throughout the entire integrated model is assured. Currently, Rome Laboratories has not begun this phase of the integration effort.

## 5. Summary

Throughout this document we have offered what we recommend as an effective outline for conducting analysis on the integrated model. Many of the individual items listed have been extensively developed in the simulation literature. However, little work has been done on the development of integration methods for these items in one simulation model. However, the integrated simulation model, perhaps because of its sheer size, demands that careful consideration be given to this issue of method integration. We believe that adherence to the items listed throughout this document will provide the *best* means of accomplishing a truly integrated simulation model at Rome Laboratories. Furthermore, we project that a successful implementation of all of the items mentioned herein would take on the order of 2-5 man-years to complete.

# **HIGH PERFORMANCE MICROSTRIP ARRAYS FOR POLARIMETRIC BISTATIC RADAR(PBR) APPLICATIONS**

**Marat Davidovitz**  
Assistant Professor  
Department of Electrical Engineering  
University of Minnesota

## **Abstract**

Novel approaches to Polarimetric Bistatic Radar antenna design were investigated with the goal of improving such parameters as cross-polarization and isolation. A feeding arrangement for linear microstrip arrays, which facilitates compact array configurations, was proposed. A design methodology was developed to implement the developed designs. A small-scale prototype was built and tested in order to verify the viability of the new concepts. Preliminary calculations were carried out to design a full-scale array for the PBR experimental facility at the RL, Hanscom AFB.

# 1 Introduction

Antennas used for polarimetric radar measurements are required to satisfy a unique combination of criteria. Among these are low cross-polarization, high isolation between orthogonal polarization channels, large azimuthal resolution and low side lobes. Moreover, mechanical considerations may present additional constraints on the size and weight of the antenna structure. This may be particularly important for bistatic measurement configurations, wherein the antennas may be re-deployed many times in the course of a measurement run.

The maturing microstrip antenna technology may afford the solution to the problems posed by polarimetric radar. Arrays of microstrip elements can provide the required polarization and beam capabilities. Microstrip designs also have the added advantages of being small(thin), light-weight, and relatively inexpensive.

The primary task described in this report concerns the design of a receiving microstrip antenna for a bistatic polarimetric radar system. The work was carried out at the Rome Laboratories ERCE section, under the auspices of the AFOSR Summer Faculty Program.

The following guidelines were issued for the design:

- Frequency: Center  $f_c = 3.200\text{GHz}$ , Band Width  $BW \approx 6\%$ .
- Polarization: Dual Orthogonal Linear (H,V)
  - Cross-Polarization  $\leq 20\text{dB}$
  - Orthogonal Port Isolation  $\geq 25\text{dB}$
- Patterns
  - 3dB Beam Width HPBW  $\leq 6^\circ$
  - Side Lobe Level SLL  $\leq -20\text{dB}$

The proposed approach entails the design of a microstrip array containing square patch elements. The array is to be fed by a hybrid microstrip/coaxial feed network incorporating an illumination taper required to achieve the specified SLL. The design tasks which have been carried out to date are:

- Design of the single element for the proper resonant frequency
- Design of a  $2 \times 2$  low-cross polarization sub-array module, which will form the basic building block for the final structure



- Design of the feed network, including the proper amplitude taper for the specified SLL
- Analysis of random phase error effects on the SLL
- Photographic mask layout of the preliminary design for the purpose of experimental verification

The listed tasks are described in more detail in the subsequent sections of the report.

## 2 Single Element Design

There exist numerous methods for analyzing the properties of single element patch antennas. They range from the analytically simple first order solutions to the rigorous, full-wave methods, requiring extensive computation. The latter, if properly executed, yield the most accurate results.

In order to obtain a successful design in as few iterations as possible, a numerical code based upon a full-wave electromagnetic analysis was utilized to find the patch dimension  $a$  for the specified resonant frequency ( $f_c = 3.20\text{GHz}$ ) and the given substrate permittivity and thickness, as well as to compute the input impedance of the patch over the frequency band of interest. The analytical details have been documented in previous publications [1],[2], [3]. The last reference contains a polynomial expression for the resonant frequency of a rectangular patch of a given dimension. The expressions are reported to be valid over a wide range of patch aspect ratios and substrate permittivities. Although these expressions are not directly useful in design, they can be inverted by means of root searching procedures. Such an algorithm was implemented in the Fortran program attached in the Appendix.

The design documented in this report was somewhat constrained by the availability of substrate material. Based upon bandwidth considerations, use of low dielectric constant, thick substrates is desirable. As clearly illustrated in Figure 1, for a fixed dielectric constant (in this case  $\epsilon_r = 2.33$ ) the bandwidth is almost a linear function of the substrate thickness. The preliminary design was carried out for a substrate characterized by  $\epsilon_r = 2.33$  and thickness  $t = 3.175\text{mm}$ , which was the most suitable material available locally. Referring back to Fig. 1 it is seen that the estimated BW for a 2:1 input SWR is approximately 4%. To achieve greater bandwidth, a substrate which either has a lower  $\epsilon_r$  or is slightly thicker needs to be utilized.

The resonant dimension of the square patch was calculated with the aid of the aforementioned Fortran code. For the chosen substrate ( $\epsilon_r = 2.33, t = 3.175\text{mm}$ ) this dimension is  $a = 27.9\text{mm}$ . This information was subsequently fed into a full-wave code to calculate the input impedance for the edge-fed square patch with  $a = 27.9\text{mm}$ . The feed point was centered between the patch corners. The results are shown in Figure 2. The displayed data are normalized with  $50\Omega$ . As expected, the impedance locus crosses the real axis at  $3.2\text{GHz}$ . The real part of the input impedance at that frequency is  $R_c = 244\Omega$ .

### 2.1 Low Cross-Polarization Four-Element Module

The single element square patch antenna ordinarily exhibits a high level of cross-polarization. This is primarily due to the presence of impurity fields associated with higher-order modes

excited by the feed line. Several means of suppressing the high-order modes have been developed in the past [4], [5]. In the case of the single element antenna, the dominant impurity modes can be eliminated by exciting the patch at the opposite sides with  $180^\circ$  difference between the two ports. Alternatively, one can implement the same idea in the form of a four single-port element sub-array module. In this configuration the four elements are paired and the two pairs are fed at the opposite sides  $180^\circ$  out of phase. In this arrangement the fields associated with the  $E_{02}$  modes, which are the primary contributors to the cross-polarization, approximately cancel each other in the far field. At the same time, the  $E_{10}$ -mode fields radiated by all four patches will reinforce each other to form the desired co-polarized field.

Shown in Figure 3 is a four patch sub-array designed in accordance with the outlined principles. This specific layout was produced for the  $\epsilon_r = 2.33$ ,  $t = 3.175$  mm substrate and the resonant frequency  $f_c = 3.2$  GHz. The feed lines between the  $50\Omega$  coaxial connector and the  $244\Omega$  microstrip patch input port are  $155\Omega$ ,  $100\Omega$  and  $70.7\Omega$ .

## 2.2 Dual Linear Polarization Array Layout

The most straightforward approach one can take to produce two orthogonal linear fields of high purity is to use two separate four element modules described in the preceding section. The two sub-arrays would have to be positioned at  $90^\circ$  with respect to one another. This arrangement is inefficient from the standpoint of substrate area utilization. In order to improve upon this design it is necessary to use the same set of four patches to transmit both polarizations. Thus each patch will be fed at two ports to produce orthogonal modes in the element 'cavity' and orthogonal polarizations in the far field. Although this is possible in principle, the actual implementation is complicated by the fact that two non-intersecting feed networks are to be positioned in a limited area around the four patches. The limits of this area are determined by the spacing of the elements in the overall array. A possible arrangement which may solve the layout problem for the linear array under consideration is shown in Figure 4. In the proposed geometry, each vertical (with respect to the line of the complete array) pair of elements forms a part of two orthogonally polarized modules. Thus constructed, the dual polarized array occupies the same substrate area as would be needed to support an array of single polarization modules.

There is a trade-off between the physical compactness of the array and the level of isolation between the orthogonal polarization input ports. Use of two completely separate four element sub-arrays to radiate the Co- and X-Pol fields yields the highest level of isolation. Reported results [6] indicate that the signal in the orthogonal port for such an arrangement is typically more than 40dB below the input level. In the more compact arrangements,

Table 1: Input Amplitude Taper

Element #	1	2	3	4	5	6	7	8
Attenuation (dB)	10	6	2	0	0	2	6	10

the proximity of the various feed lines belonging to the orthogonally polarized sub-arrays is expected to lead to stronger cross-coupling of the two input ports. A more reasonable estimate for the isolation level in this case is 25-30dB.

## 2.3 Radiation Patterns

Among the performance parameters specified for the proposed antenna are the radiation pattern Half-Power Beam Width (HPBW) and Side Lobe Level (SLL) in the azimuthal plane. Since the down range resolution is controlled by the radar, the elevation pattern can be broad. Therefore a linear array geometry would serve the purpose. To achieve the required  $\text{HPBW} = 6^\circ$  and  $\text{SLL} = 20\text{dB}$  a proper number of elements and a suitable input weight distribution have to be selected. Preliminary estimates had indicated that approximately 16 elements spaced  $d=0.6\lambda_0$  apart should produce the desired HPBW. Equivalently, 8 four element modules of the type described in the preceding sections can be used.

To achieve the required SLL a tapered element coefficients distribution is required. An initial choice for the input coefficients was based on a 30dB Chebyshev design for an 8 element array. The taper can be implemented by means of attenuators inserted into the coaxial feed lines carrying the signal from the power divider to the individual sub-arrays. The array coefficients obtained from the initial Chebyshev design corresponded to non-integral values of attenuation. Since fixed-value attenuators are typically available in steps of 1dB, it would be considerably simpler to have an input amplitude distribution which can be realized with commercially available units. Using the initial Chebyshev design as a guide, an iterative simulation procedure was carried out in order to obtain an illumination distribution quantized in steps of 1dB. The final design, summarized in Table I, yielded a theoretical radiation pattern in which the SLL did not exceed 24dB. The azimuthal patterns of a linear array of 8 four-element modules spaced  $0.6\lambda_0$  apart and excited with the proposed amplitude taper are shown in Figure 5. Both the horizontal and the vertical polarization patterns are displayed. It is clear from these plots that the worst side lobe amplitude is below 24dB for either polarization. This is 4dB lower than the specified value. This margin of safety is important in offsetting several potential error sources which were not explicitly accounted for in the design process, and which may bring the SLL up in the actual antenna. The most important of these are phase errors in the array input signals and spurious radiation from the

microstrip feeding lines. The effects of the latter are quite difficult to estimate analytically. In order to gauge the effects of random phase errors, the patterns of the array recomputed with randomly generated phase error distributions. A Gaussian random number generator was used to produce the input phase distributions. The results of simulations with various types of random phase errors are shown in Figures 6-9. The first two of these show the H- and V-Pol patterns obtained with a Gaussian phase error with zero mean and  $5^\circ$  standard deviation. The consequence of this induced error is a 2dB deterioration in the SLL, placing the worst side lobe amplitude at approximately 22dB. Therefore, this error magnitude is tolerable. However, as Figures 8,9 illustrate, when the standard deviation of the random error is increased to  $10^\circ$ , the SLL exceeds the allowed level. This leads to the conclusion that the accuracy of the fabrication and assembly procedures must be sufficient in order to ensure phase errors smaller than approximately  $5^\circ$ .

### 3 Preliminary Test Results

The modules shown in Figs. 3,4 have been fabricated and tested on an HP8510 Network Analyzer to ascertain such parameters as the resonant frequency, impedance bandwidth and in the case of the two-input-port sub-array in Fig. 4 the interport coupling. The measured frequency dependence of the various parameters is illustrated in Figures 10-13.

The input SWR for the four-element array in Fig.3 was measured in order to find the resonant frequency and the bandwidth. The network analyzer output is attached in Figure 10, whence it is seen that the measured resonant frequency is 3.215 GHz, and the 2:1 SWR bandwidth is 130 MHz. The discrepancy between the experimental and the specified value of  $f_c$  is on the order of 0.5%. The BW corresponds very closely with the theoretical predictions. The input SWR at the ports of the two module array shown in Fig. 4 behaves in the same fashion as in the isolated module case. The measured results are given in Figs. 11-12.

The interport coupling in the two-port sub-array represents a very important performance parameter. Any signal coupled from port one into port two may, if the later port is unmatched, reradiate energy into the undesired polarization. Minimization of this coupling mechanism constituted an important design consideration. The measured intercoupling data are displayed in Figure 13. At the center frequency of the design band the coupled signal is on the order of -31dB. At the band edges it is slightly higher, -25dB and -27dB, respectively. It may therefore be concluded that the executed design very effectively isolates the two orthogonal polarization ports.

Co- and X-Pol radiation patterns of the two-module sub-array were measured for both the H and the V polarizations. The patterns taken at the resonant frequency are shown in Figures 22-25. Figure 22,23 represent the Co-pol,X-pol patterns, respectively, obtained when the H-port is excited. The X-pol level for this case is 20 dB. The V-port Co- and X-pol patterns are shown in Figs. 24,25, respectively. In this case the the cross-polarized field is approximately 17db lower than the co-polarized components. This value somewhat exceeds the specified 20 dB X-pol target. The primary reason for the discrepancy is the proximity of the networks feeding the H and V ports. Power from either feed tends to couple to the other network and eventually is reradiated into the X-pol component. Several simple ways for alleviating this problem can be suggested. The simplest would require the inter-element spacing to be increased from the current value of  $0.6\lambda_0$  to  $0.7\lambda_0$  or greater, thereby loosening the coupling between the feeds. Another approach, one which would result in the least amount of coupling between the two feeds, would require the complete physical separation of the arrays for the H and V polarizations. These alternate arrangements will be evaluated experimentally in the follow-up design.

## References

- [1] M. Davidovitz, Y. T. Lo, "Rigorous Analysis of a Circular Patch Antenna Excited by a Microstrip Transmission Line," IEEE Trans. Antennas Propagat., vol. 37, no. 8, pp. 949-958, Aug. 1989.
- [2] G. Splitt, M. Davidovitz, "Guidelines for Design of Electromagnetically Coupled Microstrip Patch Antennas on Two-layer Substrates," IEEE Trans. Antennas Propagat., vol. 38, no. 7, July 1990.
- [3] W. C. Chew and Q. Liu, "Resonance Frequency of a Rectangular Patch", IEEE Trans. Antennas Propagat., vol. AP-36, Aug. 1988.
- [4] T. Chiba, "Suppression of Higher Order Modes and Cross-Polarized Component for Microstrip Antennas", IEEE AP-S Symp. Digest, May 1982, pp. 208-288.
- [5] J. Huang, "Low Cross-Polarization Linearly Polarized Microstrip Array", IEEE AP-S Symp. Digest, May 1990, pp.1750-1753.
- [6] E. Levine and S. Shtrikman, "Experimental Comparison Between Dual-Polarized Microstrip Antennas", Microwave and Guided Wave Letters, vol 3, no. 1, Jan. 1990.

## 4 Acknowledgement

I wish to acknowledge the assistance received from Dr. K.V.N. Rao, Dr. John Lennon and Mr. William Stevens. I am especially greatfull to Mrs. Michele Champion for taking the time to discuss my designs and helping me implement them, and Dr. J. Herd for his assistance and many helpful suggestions.

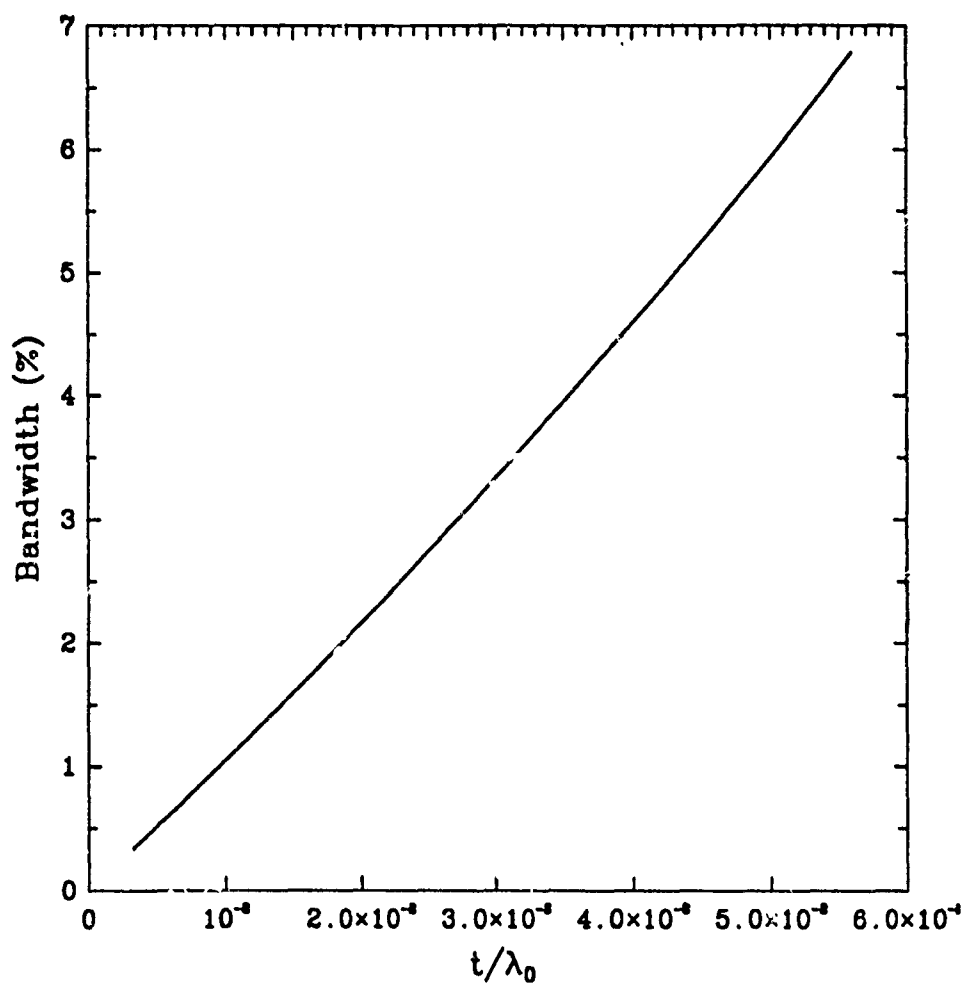


Figure 1: 2:1 SWR bandwidth versus substrate thickness for  $\epsilon_r = 2.33$ .



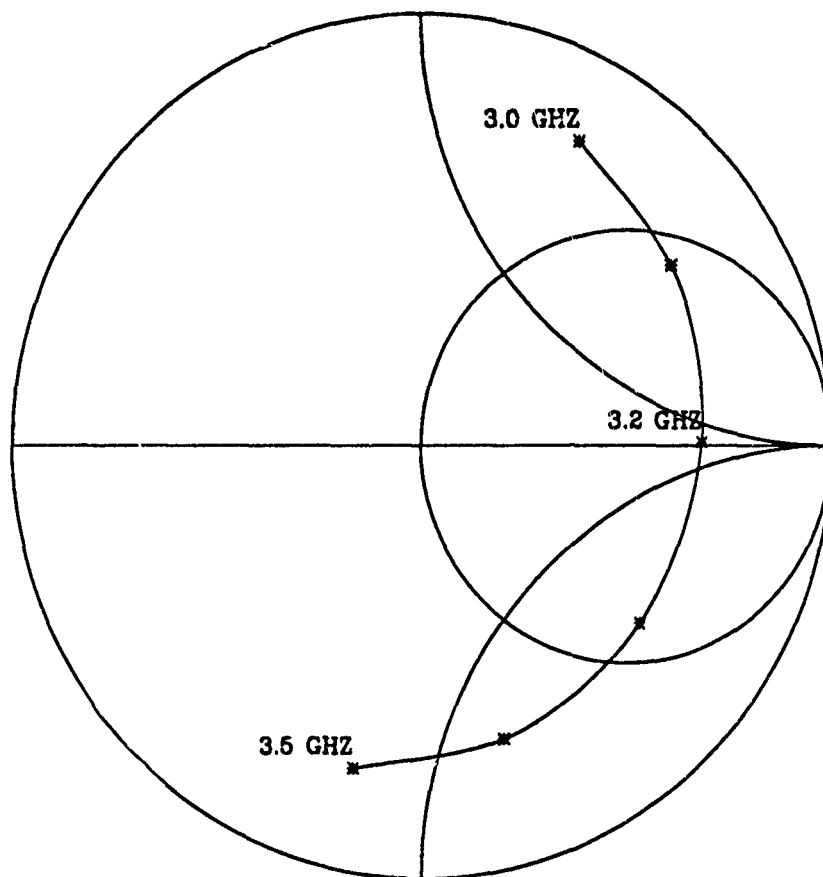


Figure 2: Calculated input impedance for a square patch antenna:  $a=27.9\text{mm}$ ,  $\epsilon_r = 2.33$ .

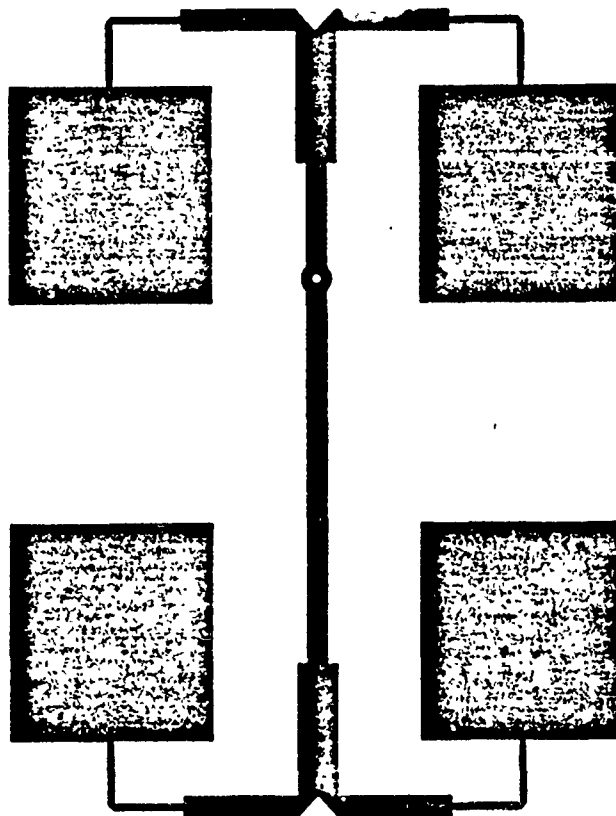


Figure 3: Four-element low cross-polarization module.

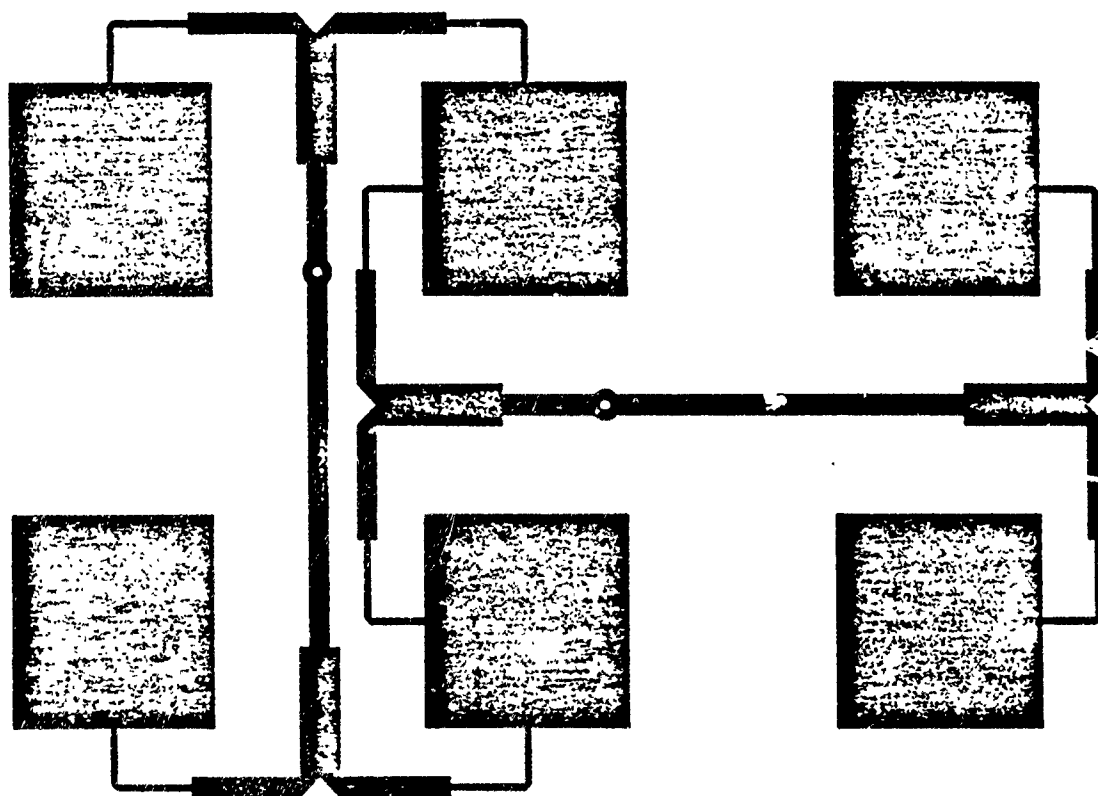


Figure 4: A compact feeding arrangement for dual polarized array.

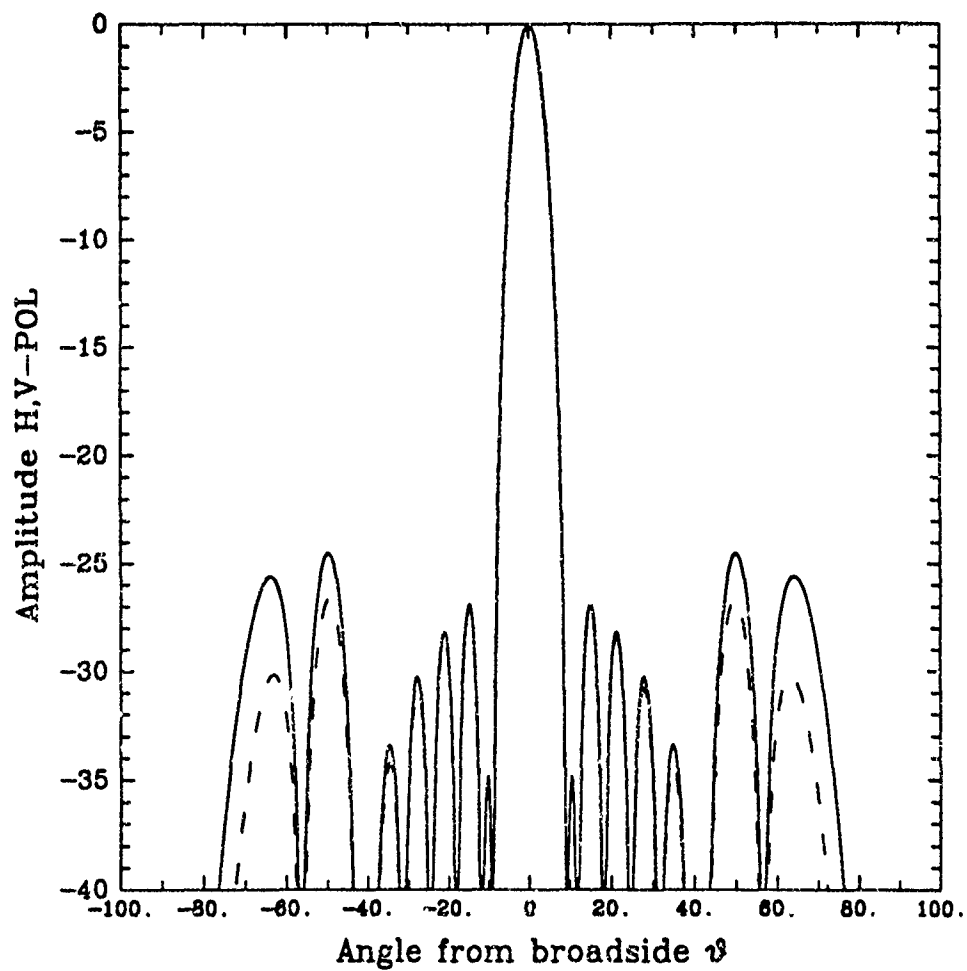


Figure 5: Horizontal and Vertical polarization power patterns.  
 (—H-POL, - - - - V-POL ;  $d=0.6\lambda_0$  )

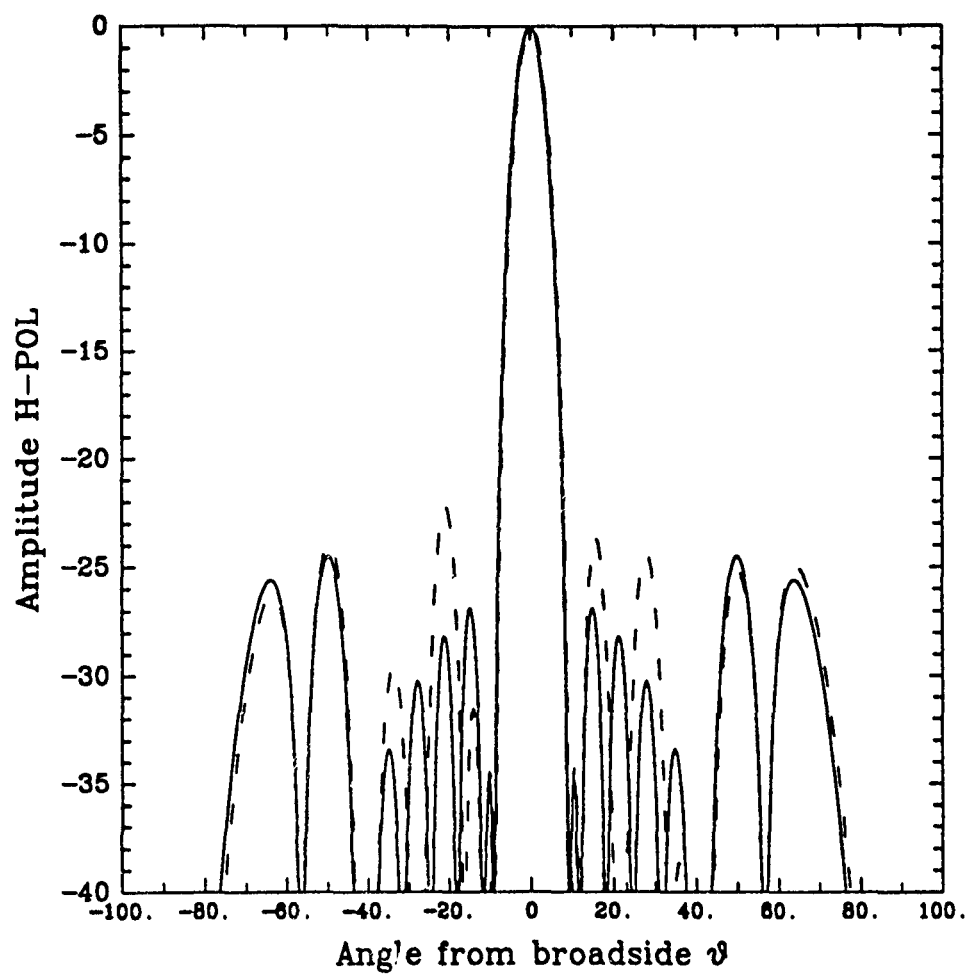


Figure 6: H-POL patterns: ——— no phase error, - - - - Gaussian phase error with  $0^\circ$  mean,  $5^\circ$  standard deviation.

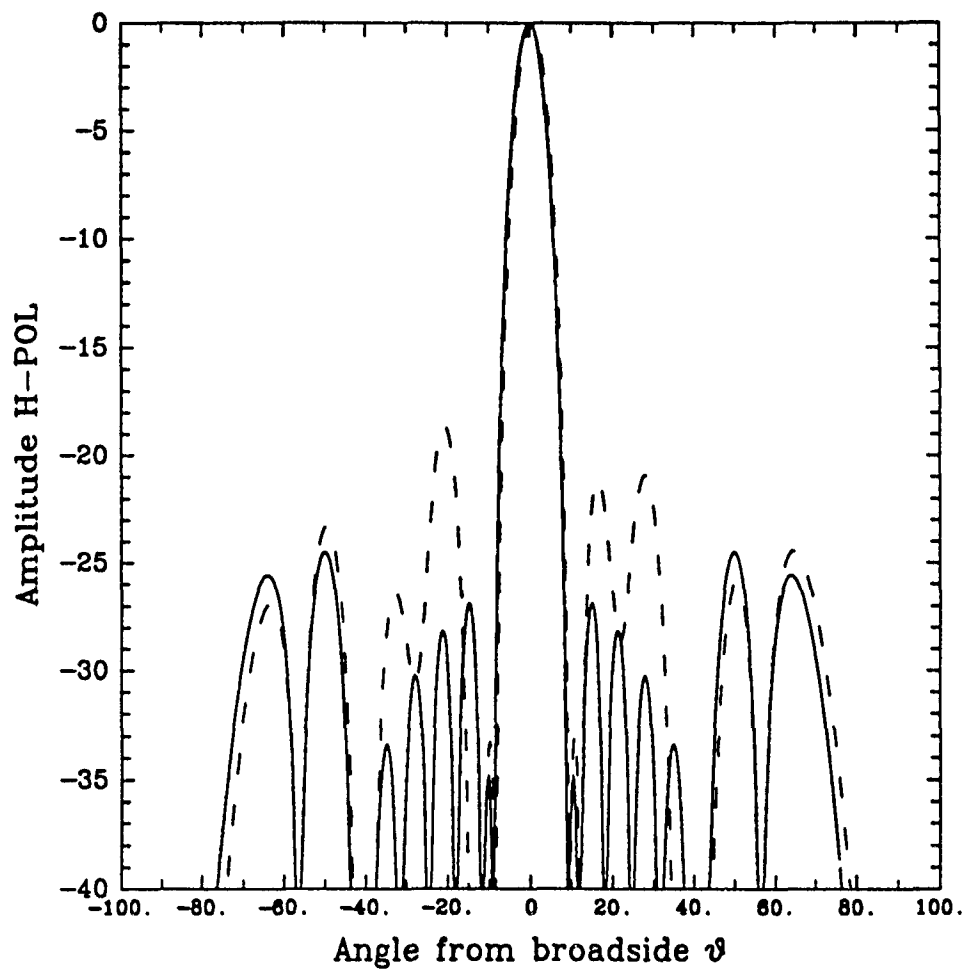


Figure 7: H-POL patterns: ——— no phase error, - - - - Gaussian phase error with  $0^\circ$  mean,  $10^\circ$  standard deviation.

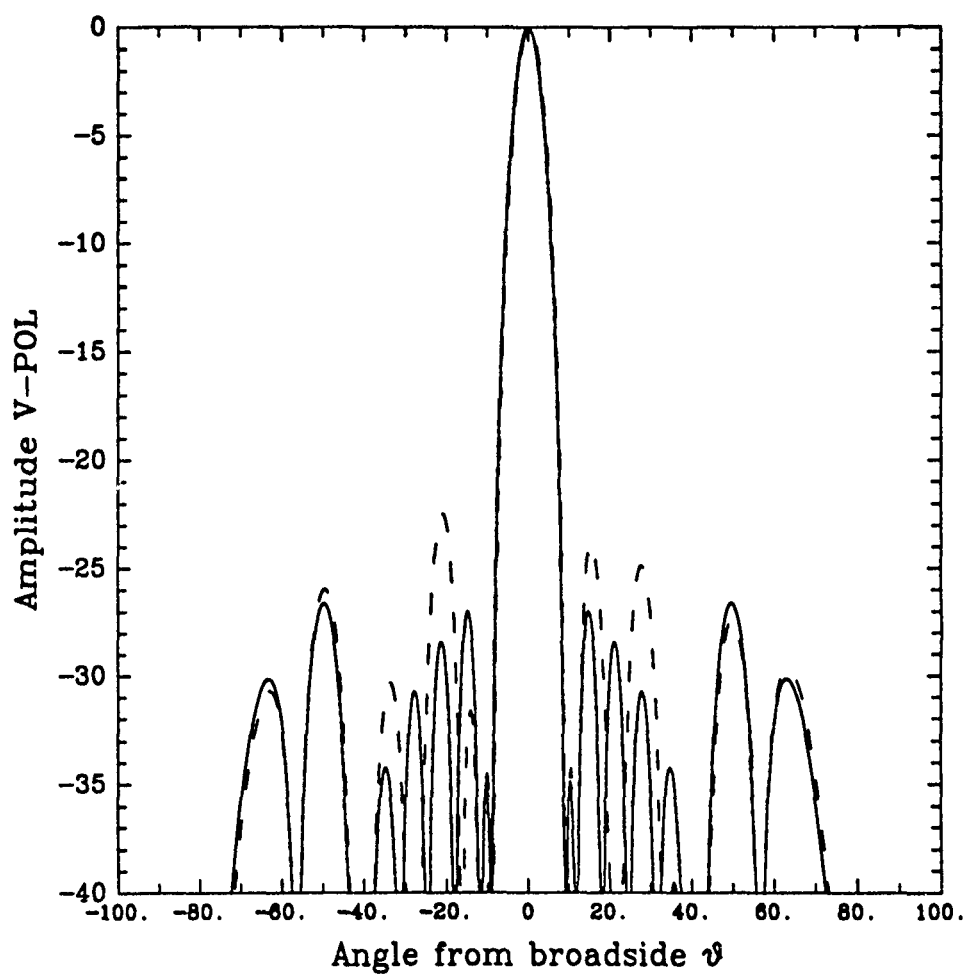


Figure 8: V-POL patterns: — no phase error, - - - - Gaussian phase error with  $0^\circ$  mean,  $5^\circ$  standard deviation.

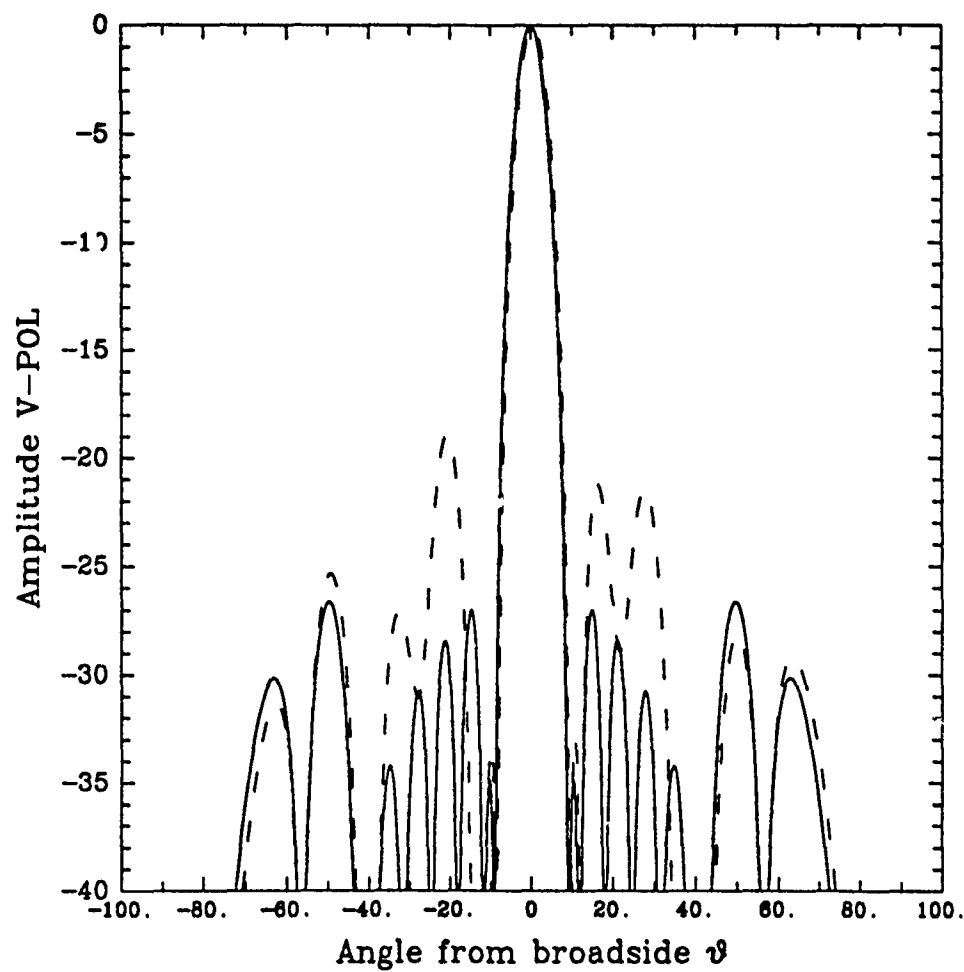


Figure 9: V-POL patterns: ——— no phase error, - - - - Gaussian phase error with  $0^\circ$  mean,  $10^\circ$  standard deviation.



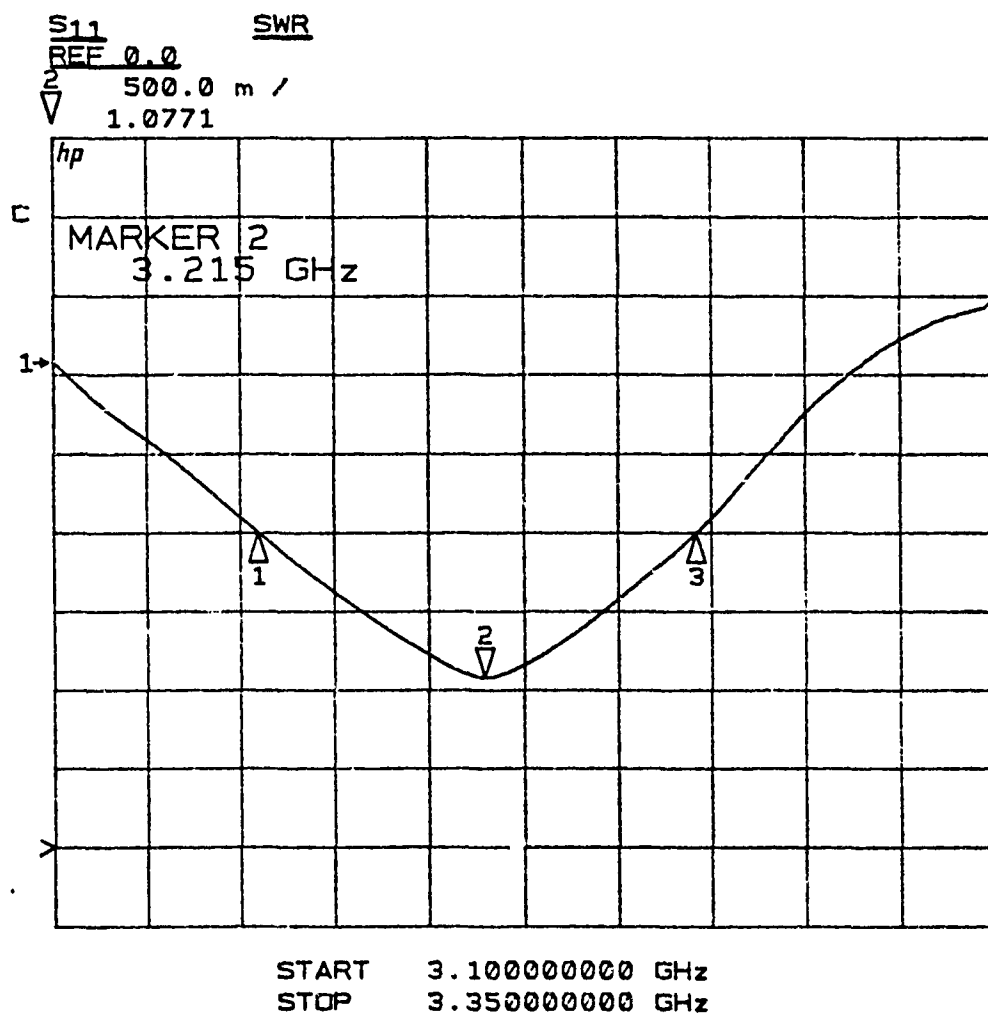


Figure 10: Input SWR for the isolated 4-element module.

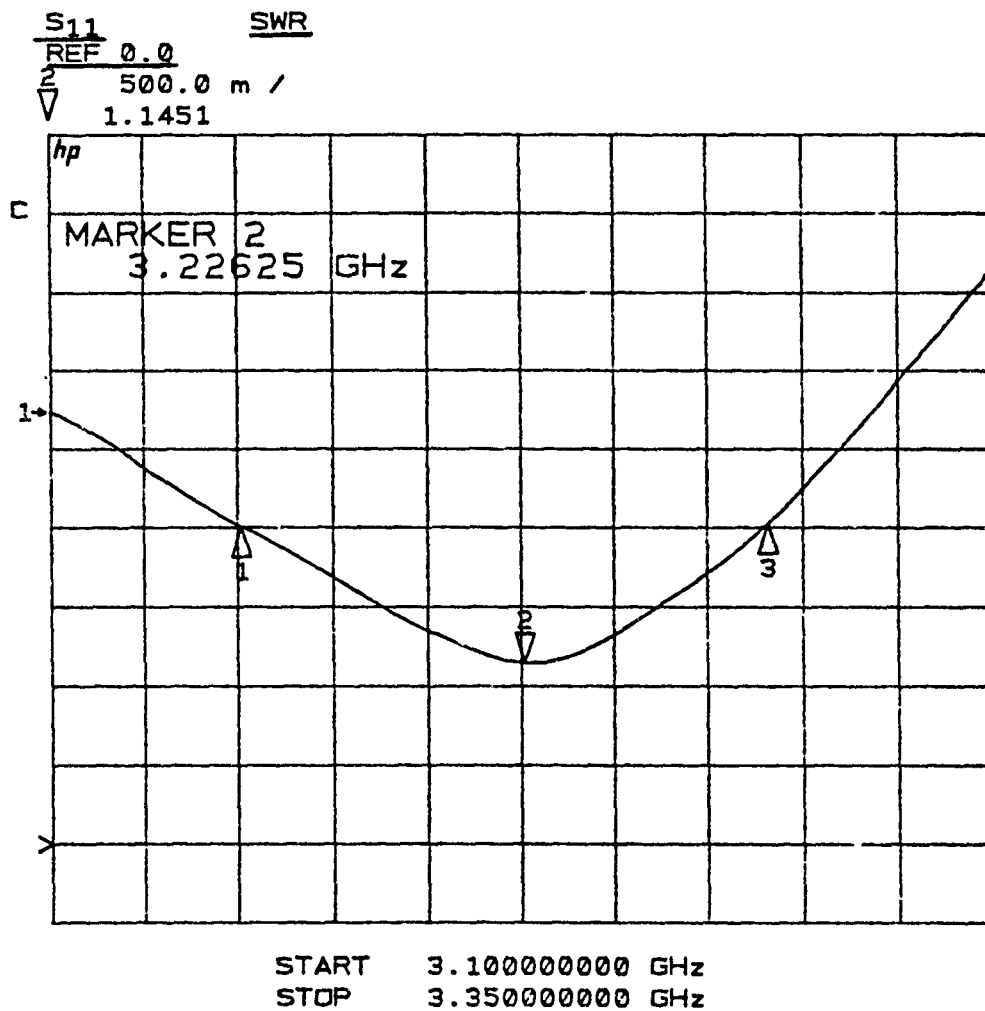


Figure 11: Input SWR for the H-Pol port of the two-module array.

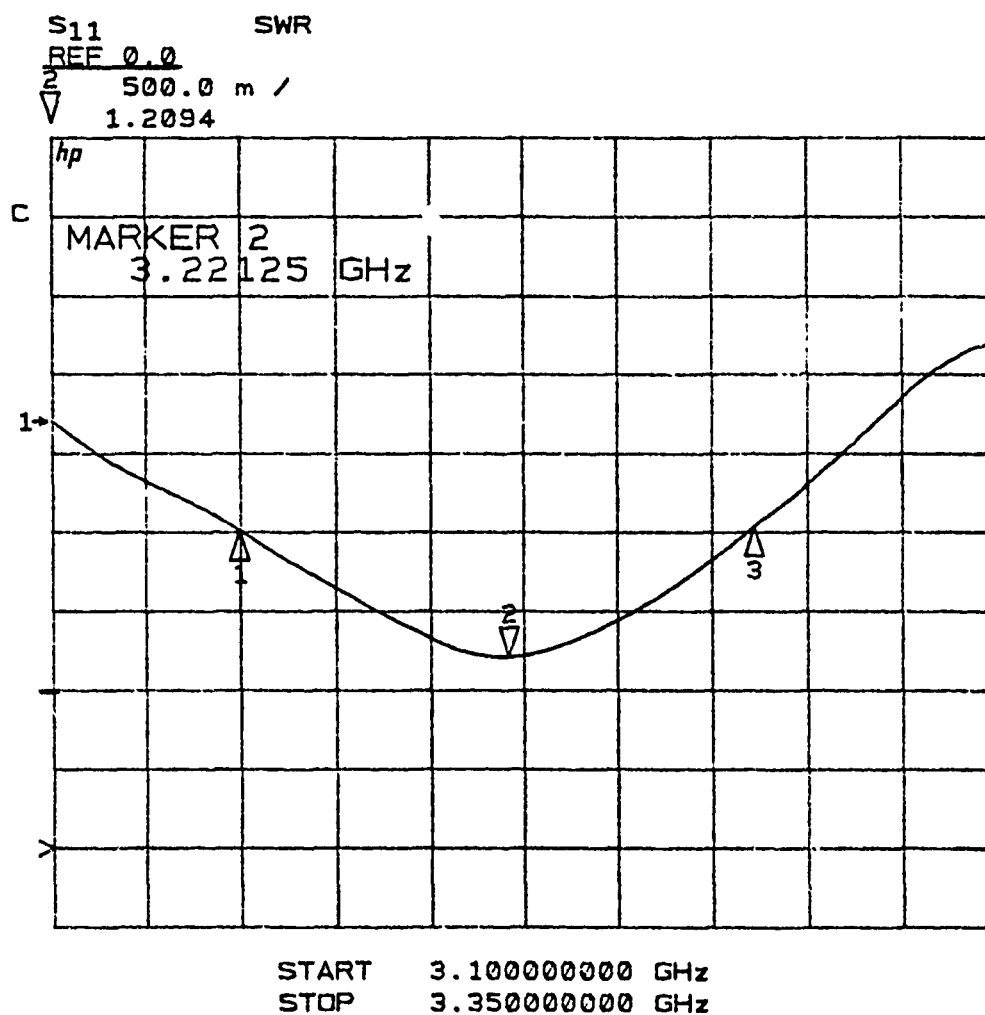


Figure 12: Input SWR for the V-Pol port of the two-module array.

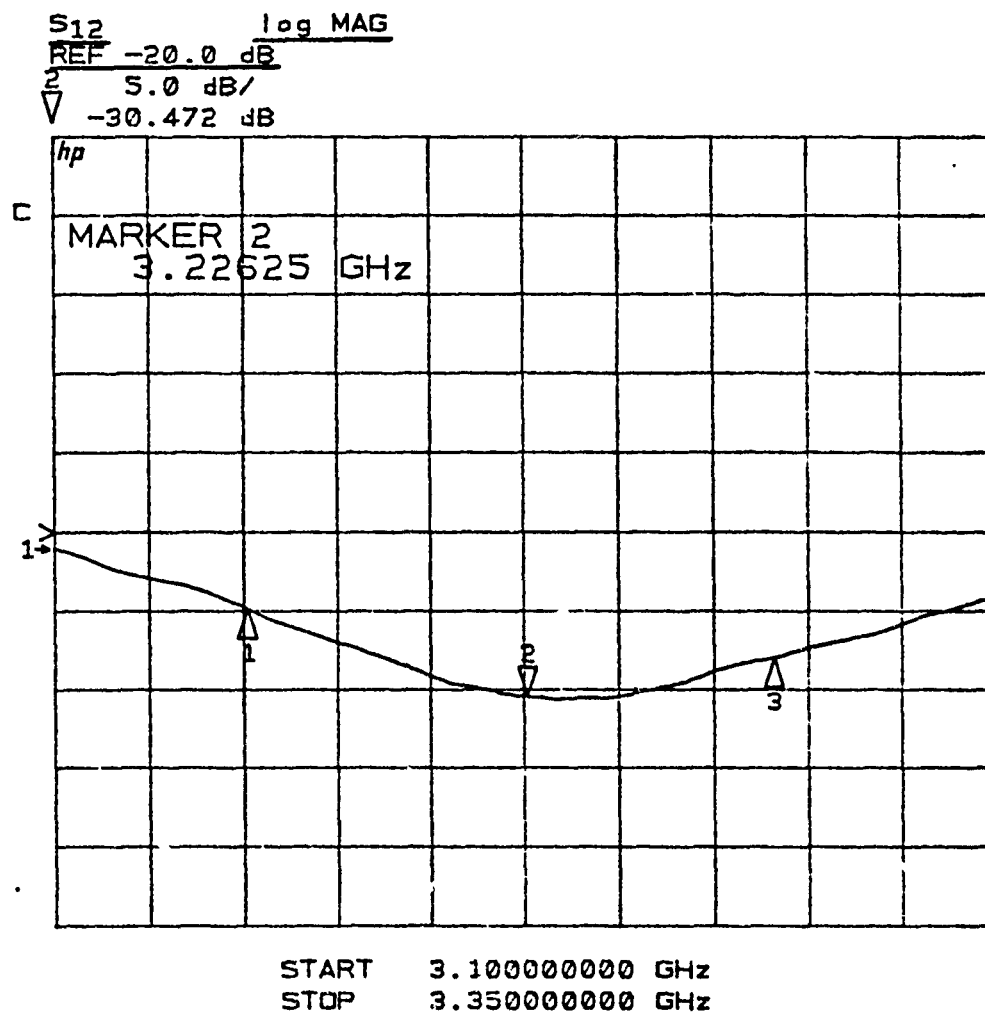


Figure 13: Interport coupling in the two-module array.

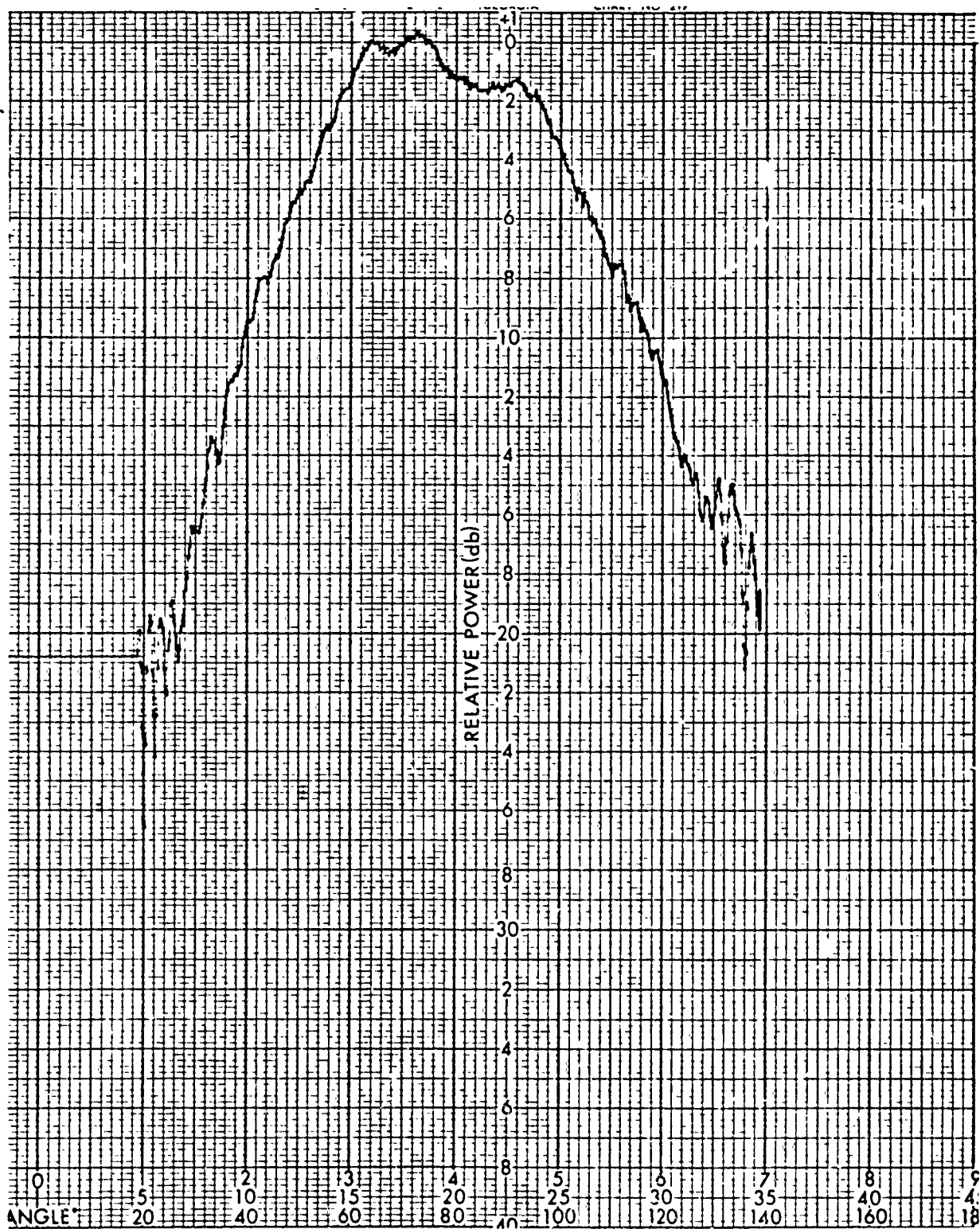


Figure 14: Horizontal co-polarized radiation pattern.

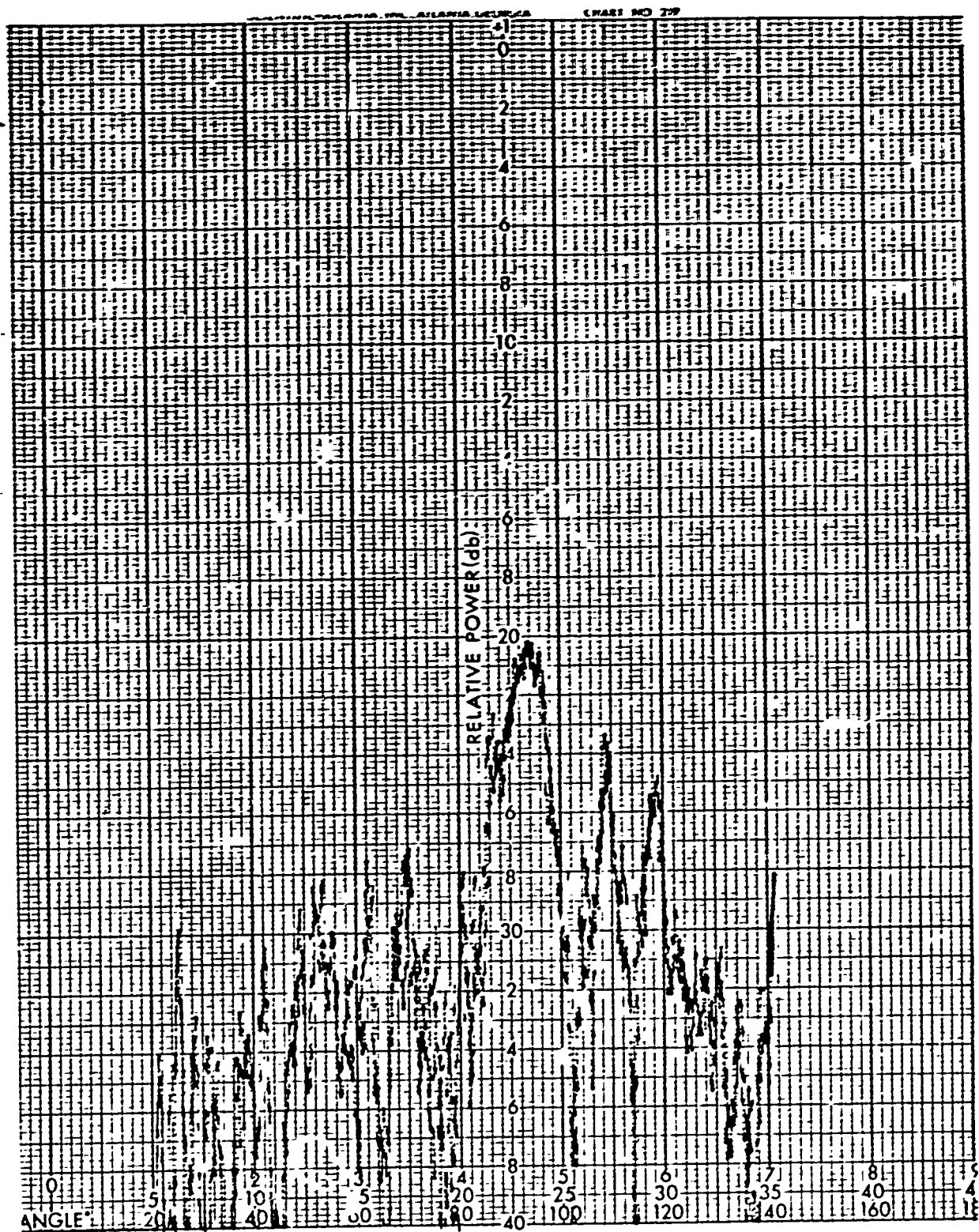


Figure 15: Horizontal cross-polarized radiation pattern.

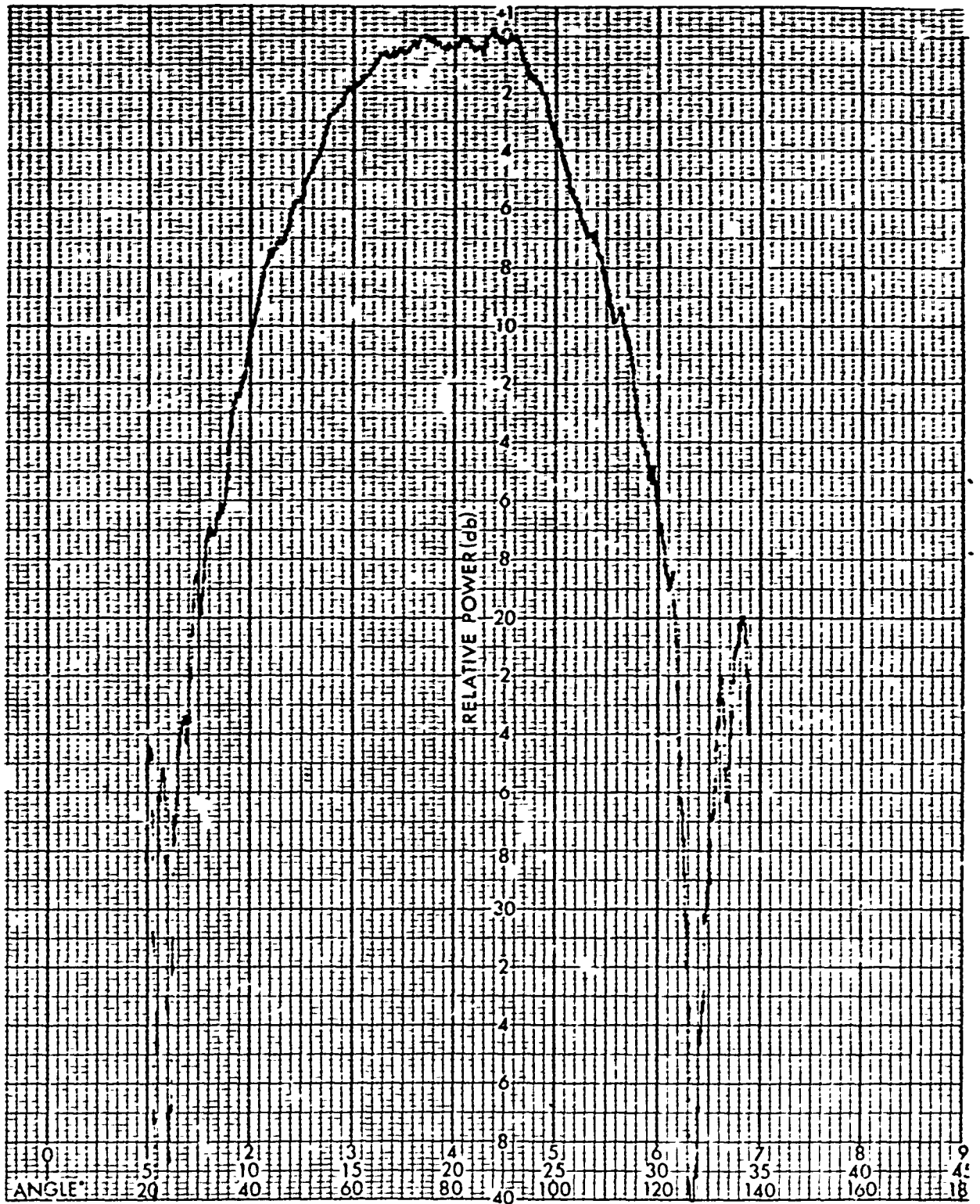


Figure 16: Vertical co-polarized radiation pattern.

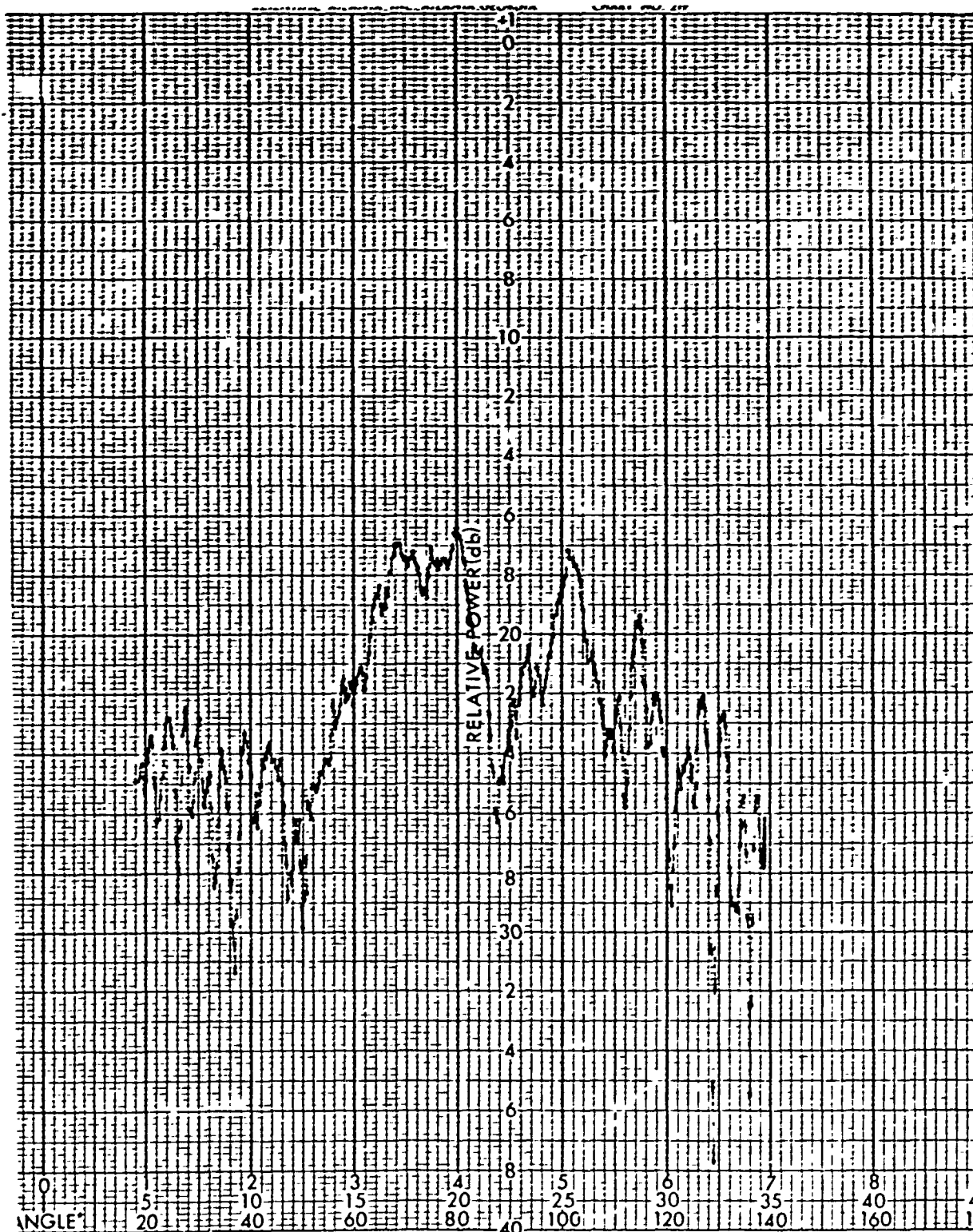


Figure 17: Vertical cross-polarized radiation pattern.





```
ldivid=lamda0/4./sqrt(epsefd)
wdivid=wohdiv*hsub
```

C CALCULATE THE FEED POINT OFFSET LAMBDA(100ohm)/4.

```
call mspara (zline,woh100,eps100)
offset=lamda0/4./sqrt(eps100)
wline=woh100*hsub
```

```
open(unit=31,file='modata')
write(31,*) 'Square Patch Size - a=',a
write(31,*) 'Quarter-Wave X-former'
write(31,*) 'Z0=',ztrans
write(31,*) 'Length=',ltrans,' Width=',wtrans
write(31,*) 'Power Divider'
write(31,*) 'Z0=',zdivid
write(31,*) 'Length=',ldivid,' Width=',wdivid
write(31,*) 'Main Feed Line'
write(31,*) 'Z0=',zline
write(31,*) 'wline=',wline
write(31,*) 'Feed Offset =',offset
close(31)
end
```

```
subroutine patsiz(freq,a)
implicit double precision (a-h,o-z)
```

```
common/fdesir/freq0
common/releps/epsr
common/thick/hsub
external resfre
```

```
freq0=freq
```

```
alow=5.*hsub
ahigh=1000.*hsub
toler=1.0d-10
a=zeroin(alow,ahigh,resfre,toler)
```

```
end
```

```
double precision function resfre(a)
implicit double precision (a-h,o-z)
```

```
common/fdesir/freq
common/releps/epsr
common/thick/hsub
```

```
call resrec(epsr,a,a,hsub,fres,q)
resfre=fres-freq
```

```
end
```

```
subroutine resrec(er,a,b,t,fres,q)
implicit double precision (a-h,o-z)
```

```
data a100,a101,a102,a103,a110,a111,a112,a113,
& a120,a121,a122,a123,a200,a201,a202,a203,
& a210,a211,a212,a213,a220,a221,a222,a223/
& -3.156754397448,1.795725629752,-0.2652696640599,
& 1.3023599874332d-2,
& 0.3030205876506,-0.8132378043906,0.1377825867855,
& -7.4426027077389d-3,
& -6.999192346233d-2,0.2074387779152,-3.7846975461942d-2.
```

```

& 2.1295393807866d-3,
& 14.47374793072,-9.117720788346,1.204181550099,
& -5.3727586612667d-2,
& -3.919355911932,4.283895471007,-0.5480065590744,
& 2.3856532089789d-2,
& 1.2689142466212,-1.322922648767,0.1932147231985,
& -9.2095501012563d-3/
data a300,a301,a302,a303,a310,a311,a312,a313,
& a320,a321,a322,a323/
& -24.64990432442,14.07938052452,-1.241188231639,
& 3.0849725458221d-2,
& 7.104107341725,-5.734060839448,3.00209083841d-2,
& 2.9052167395763d-2,
& -3.089017025567,2.374351824763,-0.176778025576,
& 9.907781784825d-4/
data b100,b101,b102,b103,b110,b111,b112,b113,
& b120,b121,b122,b123,b200,b201,b202,b203,
& b210,b211,b212,b213,b220,b221,b222,b223/
& -4.6152911d-2,2.7101483d-2,-4.4631474d-3,
& 2.1826716d-4,-0.7032761,0.2455350,-3.200742d-2,
& 1.4145372d-3,0.1795260,-7.5435149d-2,1.0676009d-2,
& -4.9243571d-4,-0.8322784,-6.4469166d-2,3.5479948d-2,
& -2.2502933d-3,3.735017,-1.411132,0.1898225,
& -8.5380571d-3,-1.135043,0.4898767,-7.0118777d-2,
& 3.2546222d-3/
data b300,b301,b302,b303,b310,b311,b312,b313,
& b320,b321,b322,b323/
& 2.135107,-0.2780477,-1.4273385d-2,2.2362657d-3,
& -7.907233,3.316067,-0.4600076,2.0703688d-2,
& 2.526486,-1.141029,0.1641063,-7.5149969d-3/

```

```

a10=a100+a101*Er+a102*Er**2+a103*Er**3
a11=a110+a111*Er+a112*Er**2+a113*Er**3
a12=a120+a121*Er+a122*Er**2+a123*Er**3
a20=a200+a201*Er+a202*Er**2+a203*Er**3
a21=a210+a211*Er+a212*Er**2+a213*Er**3
a22=a220+a221*Er+a222*Er**2+a223*Er**3
a30=a300+a301*Er+a302*Er**2+a303*Er**3
a31=a310+a311*Er+a312*Er**2+a313*Er**3
a32=a320+a321*Er+a322*Er**2+a323*Er**3

```

```

a1=a10+a11*(a/b)+a12*(a/b)**2
a2=a20+a21*(a/b)+a22*(a/b)**2
a3=a30+a31*(a/b)+a32*(a/b)**2

```

```

Wr=1.+a1*(t/b)+a2*(t/b)**2+a3*(t/b)**3

```

```

b10=b100+b101*Er+b102*Er**2+b103*Er**3
b11=b110+b111*Er+b112*Er**2+b113*Er**3
b12=b120+b121*Er+b122*Er**2+b123*Er**3
b20=b200+b201*Er+b202*Er**2+b203*Er**3
b21=b210+b211*Er+b212*Er**2+b213*Er**3
b22=b220+b221*Er+b222*Er**2+b223*Er**3
b30=b300+b301*Er+b302*Er**2+b303*Er**3
b31=b310+b311*Er+b312*Er**2+b313*Er**3
b32=b320+b321*Er+b322*Er**2+b323*Er**3

```

```

b1=b10+b11*(a/b)+b12*(a/b)**2
b2=b20+b21*(a/b)+b22*(a/b)**2
b3=b30+b31*(a/b)+b32*(a/b)**2

```

```

Wi=b1*(t/b)+b2*(t/b)**2+b3*(t/b)**3

```

```

w0=3.d8/(2.*a)/sqrt(Er)
fres=Wr*w0
q=-Wr/2./Wi

*   write(*,*) 'Reson. Freq.=' ,Wr*w0

end

subroutine mspara (z1,woh,epsef)
implicit double precision (a-h,o-z)

common/desirz/z0
external zchar0

z0=z1

wohlow=0.000001
wohigh=1000.
toler=1.0d-10
woh= zeroin(wohlow,wohigh,zchar0,toler)
epsef=epseff(woh)

*   write(*,*) 'w/h=',woh,' epseff=',epsef

end

double precision function zchar0(woverh)
implicit double precision (a-h,o-z)
common/desirz/z1

zchar0=zchar(woverh)-z1

return
end

double precision function zchar(woverh)
implicit double precision (a-h,o-z)
common/releps/epsr

pi=3.1415926535

f1=6.+(2.*pi-6.)*exp(-(30.666/woverh)**0.7528)
zhomog=60.*dlog(f1/woverh+sqrt(1.+(2./woverh)**2))
zchar=zhomog/sqrt(epseff(woverh))

return
end

double precision function epseff(woverh)
implicit double precision (a-h,o-z)
common/releps/epsr

a=1.+dlog((woverh**4+(woverh/52.)**2)/(woverh**4+0.432))/49.+
&      dlog(1.+(woverh/18.1)**3)/18.7
b=0.564*(((epsr-0.9)/(epsr+3.))**0.053)
epseff=(epsr+1.)/2.+(epsr-1.)/2./(1.+10./woverh)**(a*b)

return
end

c To get dlmach, mail netlib
c   send dlmach from core

```

```

double precision function zeroin(ax,bx,f,tol)
double precision ax,bx,f,tol

c
c      a zero of the function f(x) is computed in the interval ax,bx .
c
c input..
c
c ax      left endpoint of initial interval
c bx      right endpoint of initial interval
c f       function subprogram which evaluates f(x) for any x in
c         the interval ax,bx
c tol     desired length of the interval of uncertainty of the
c         final result (.ge.0.)
c
c output..
c
c zeroin abscissa approximating a zero of f in the interval ax,bx
c
c      it is assumed that f(ax) and f(bx) have opposite signs
c      this is checked, and an error message is printed if this is not
c      satisfied. zeroin returns a zero x in the given interval
c      ax,bx to within a tolerance 4*macheps*abs(x)+tol, where macheps is
c      the relative machine precision defined as the smallest representable
c      number such that 1.+macheps .gt. 1.
c      this function subprogram is a slightly modified translation of
c      the algol 60 procedure zero given in richard brent, algorithms for
c      minimization without derivatives, prentice-hall, inc. (1973).
c
      double precision a,b,c,d,e,eps,fa,fb,fc,toll,xm,p,q,r,s
      double precision dabs, dlmach
10  eps = dlmach(4)
      toll = eps+1.0d0
c
      a=ax
      b=bx
      fa=f(a)
      fb=f(b)
c
c      check that f(ax) and f(bx) have different signs
      if (fa .eq. 0.0d0 .or. fb .eq. 0.0d0) go to 20
      if (fa * (fb/dabs(fb)) .le. 0.0d0) go to 20
      write(6,2500)
2500  format(1x,'f(ax) and f(bx) do not have different signs,',
1      ' zeroin is aborting')
      return
20  c=a
      fc=fa
      d=b-a
      e=d
30  if (dabs(fc).ge.dabs(fb)) go to 40
      a=b
      b=c
      c=a
      fa=fb
      fb=fc
      fc=fa
40  toll=2.0d0*eps*dabs(b)+0.5d0*toll
      xm = 0.5d0*(c-b)
      if ((dabs(xm).le.toll).or.(fb.eq.0.0d0)) go to 150
c
c see if a bisection is forced
c
      if ((dabs(e).ge.toll).and.(dabs(fa).gt.dabs(fb))) go to 50
      d=xm
      e=a
      go to 110
c

```

```

        if (a.ne.c) go to 60
c
c linear interpolation
c
        p=2.0d0*xm*s
        q=1.0d0-s
        go to 70
c
c inverse quadratic interpolation
c
        60 q=fa/fc
           r=fb/fc
           p=s*(2.0d0*xm*q*(q-r)-(b-a)*(r-1.0d0))
           q=(q-1.0d0)*(r-1.0d0)*(s-1.0d0)
        70 if (p.le.0.0d0) go to 80
           q=-q
           go to 90
        80 p=-p
        90 s=e
           e=d
           if (((2.0d0*p).ge.(3.0d0*xm*q-dabs(tol1*q))).or.(p.ge.
           *dabs(0.5d0*s*q))) go to 100
           d=p/q
           go to 110
        100 d=xm
           e=d
        110 a=b
           fa=fb
           if (dabs(d).le.tol1) go to 120
           b=b+d
           go to 140
        120 if (xm.le.0.0d0) go to 130
           b=b+tol1
           go to 140
        130 b=b-tol1
        140 fb=f(b)
           if ((fb*(fc/dabs(fc))).gt.0.0d0) go to 20
           go to 30
        150 zeroin=b
           return
           end

```

```

        DOUBLE PRECISION FUNCTION D1MACH(I)
C
C DOUBLE-PRECISION MACHINE CONSTANTS
C
C D1MACH( 1) = B**(EMIN-1), THE SMALLEST POSITIVE MAGNITUDE.
C
C D1MACH( 2) = B**EMAX*(1 - B**(-T)), THE LARGEST MAGNITUDE.
C
C D1MACH( 3) = B**(-T), THE SMALLEST RELATIVE SPACING.
C
C D1MACH( 4) = B**(1-T), THE LARGEST RELATIVE SPACING.
C
C D1MACH( 5) = LOG10(B)
C
C TO ALTER THIS FUNCTION FOR A PARTICULAR ENVIRONMENT,
C THE DESIRED SET OF DATA STATEMENTS SHOULD BE ACTIVATED BY
C REMOVING THE C FROM COLUMN 1.
C ON RARE MACHINES A STATIC STATEMENT MAY NEED TO BE ADDED.
C (BUT PROBABLY MORE SYSTEMS PROHIBIT IT THAN REQUIRE IT.)
C
C FOR IEEE-ARITHMETIC MACHINES (BINARY STANDARD), ONE OF THE FIRST
C TWO SETS OF CONSTANTS BELOW SHOULD BE APPROPRIATE.
C
C WHERE POSSIBLE. DECIMAL, OCTAL OR HEXADECIMAL CONSTANTS ARE USED

```

# USER ASSISTED INFORMATION EXTRACTION

Pradip Peter Dey

Hampton University

Computer Science Department

Hampton, VA 23668

## 0. ABSTRACT

User assisted information extraction from formatted and unformatted messages requires a friendly user interface, knowledge based natural language processing, and utilities for integrating future developments in message processing technology. Parallel and distributed processing will play an important role in the future enhancement of the message processing technology. Two research tools were developed in the summer of 1991 that allow one to experiment with a prototype information extraction system available in the IRDS branch of Rome Laboratory. Some components of the prototype system can be executed concurrently by means of these tools. In order to exploit the enormous potential of the system a detailed study of parallel processing architectures, algorithms, data and control structures for the system is recommended.

## 1. INTRODUCTION

This report describes the research conducted between June 3 and August 9, 1991 in the IRDS branch of Rome Laboratory under the AFOSR program. The research deals with user assisted information extraction. One of the goals of IRDS has been to develop and test a generic system in this area. A prototype system called Generic Intelligence Processor (GIP) is available in IRDS for experimental purposes. An enhanced version of GIP is under development now. In addition, further long-term enhancements of GIP are already being planned. With GIP, IRDS has undertaken a challenging task of putting an emerging technology into work. GIP is proposed to be developed in evolutionary stages gradually adding new components and enhancing the system as the technology develops. GIP has enormous potential for technology transfer in some crucial areas as it provides an environment for transferring both short and long-term developments. This research focuses on long-term enhancement of GIP rather than immediate enhancement that is already undertaken. The specific long-term enhancement that is addressed here is the parallelization problem of GIP. GIP must be studied for parallel and distributed execution for two main reasons: (a) GIP is a complicated system requiring substantial computing power, (b) Since



GIP would be used in a distributed environment, resource and load sharing among the processors of the environment is a reasonable goal.

Two research tools for investigating parallel and distributed execution of GIP modules have been developed during the summer of 1991. The tools are: (1) Remoteacquire, and (2) Remotenlp. Remoteacquire is used for acquiring a large number of messages from a remote source. Remotenlp is used to process several messages on a remote site and obtain the results back to the host. These tools are described in section-3. These tools allow Rome Laboratory to conduct some preliminary experiments in parallel execution of GIP modules. A detailed study of parallel and distributed execution of GIP has to be conducted at a later time.

## 2. GENERIC INTELLIGENCE PROCESSOR (GIP)

A detailed description of GIP is available in a technical proposal written by the Knowledge Systems Concepts, Inc. (KSC 1991). The main goal of GIP is to extract information from formatted and unformatted messages and transmit the results to downstream processes such as expert systems. GIP is a toolkit for developing, testing, and delivering message processing applications. When fully developed GIP will provide the following capabilities (a) user selectable input sources, (b) automatic message parsing, (c) automatic/user assisted information extraction, (d) automatic output formatting for development systems, (e) utilities for developing new algorithms. For processing messages with substantial natural language texts user assistance will be crucial for GIP, because, natural language processing technology is not mature enough for fully automated information extraction in an operational environment. Therefore, there is considerable emphasis on user interface in the GIP design. The GIP program components are reproduced in Figure-1.

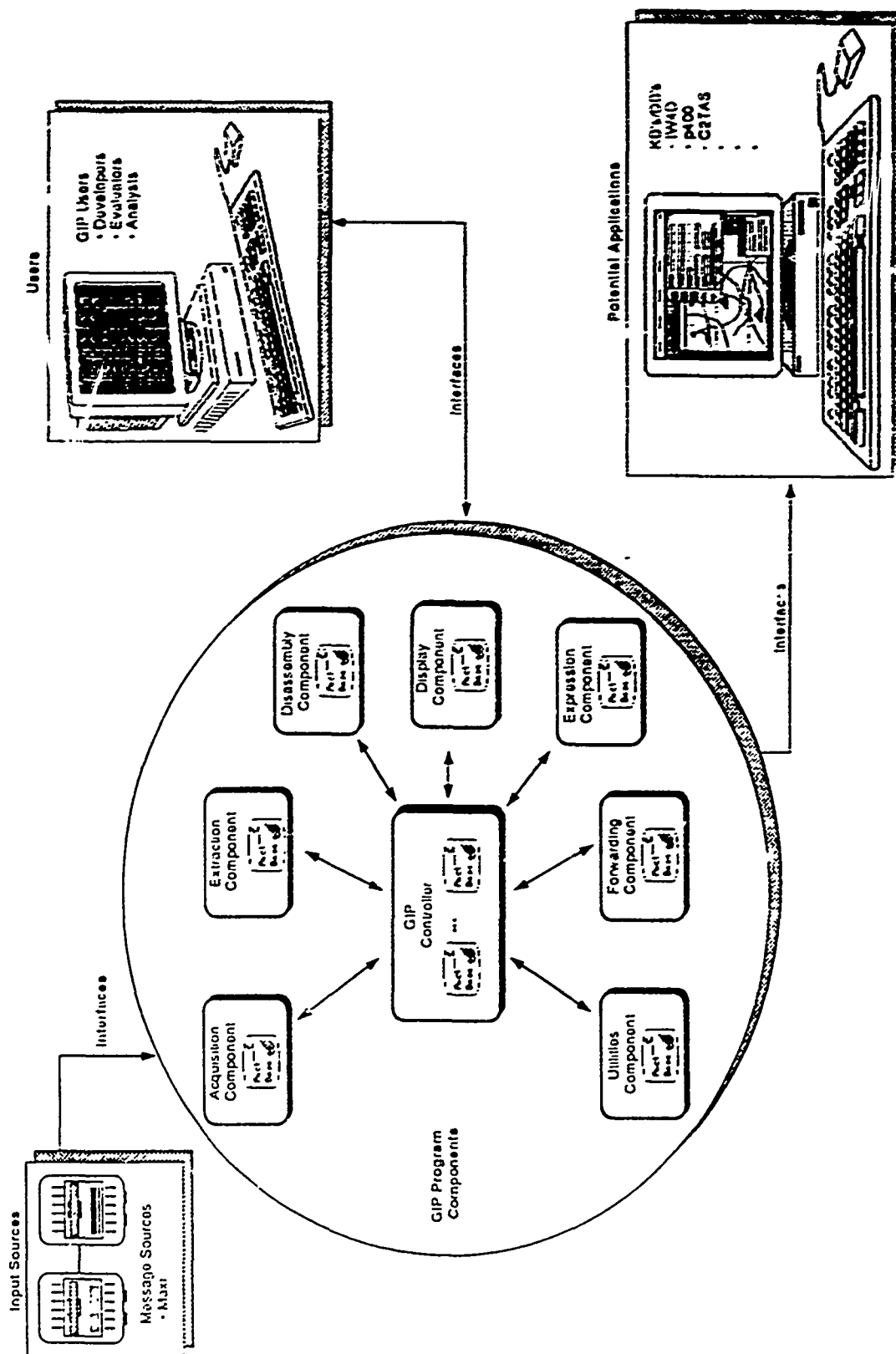


Figure-1: GIP Program Components (from KSC 1991)

As mentioned earlier, GIP is a complicated system requiring enormous computing power. Efficient execution of GIP is possible only by clever utilization of computing resources in the GIP environment. In order to achieve this, parallel and distributed execution of GIP modules would be necessary. A detailed study of parallel processing architectures, algorithms, data and control structures for GIP will be required. This preliminary study provides two tools that allow user controlled concurrent execution of some GIP components for experimental purposes.

### 3. TWO EXPERIMENTAL TOOLS

GIP must perform in a distributed environment where several computers are networked together for sharing resources and computation. Two tools have been developed that allow this sharing on an experimental basis. These tools are also useful to analyze GIP's parallel processing requirements and design and development. The tools are called Remoteacquire and Remotenlp.

#### 3.1. Remoteacquire

Remoteacquire is a tool that acquires messages from remote sources. It uses the rcp protocol from Networking Tools and Programs (Sun Microsystems) for copying an entire directory from a remote source. It prompts for the name of a remote source from which messages have to be acquired. Once the name is given, the messages are acquired over the network which can be processed by GIP's other modules. Remoteacquire can be concurrently executed with some other components (such as the extraction component) of GIP.

### 3.2. Remotenlp

Remotenlp is a tool that executes a natural language processor (NLP) on a remote site and obtains the results back in the GIP environment. It uses the rsh protocol from Networking Tools and Programs (Sun Microsystems). While the NLP is being executed on the remote site, GIP's other modules can be executed on the local host. This produces large grain parallelism at a high level. Parallelism at another level is simulated by Remotenlp by creating several processes and executing a different message in each of the processes.

## 4. PARALLEL ARCHITECTURES FOR GIP

There are three basic architecture types: (1) SISD = single instruction single data stream. Traditional sequential computers are of this type. A single instruction is executed at a time on a single piece of data. (2) SIMD = single instruction multiple data stream. A single instruction is executed in a number of processors on different pieces of data. Synchronization is automatic in this architecture. A processor either executes the same instruction with other processors or remains idle. The connection machine is an SIMD machine (Hillis 1985). (3) MIMD = multiple instruction multiple data stream. Different instructions can be executed in different processors on different pieces of data. There are two types of MIMD computers: (a) Multicomputers = loosely coupled systems. In this architecture each processor has its own local memory and communications between the processors are achieved through an interconnection network. They are also known as distributed systems. The hypercube computers are of this type (Athas and Seitz 1988).

(b) **Multiprocessors** = tightly coupled systems. In these systems, the memory is common to all processors. The processors share a number of global variables stored in the common memory. These variables enable the processors to communicate efficiently through the shared memory if the number of processors is not too large. Two examples of commercially available shared memory multiprocessors are the BBN Butterfly system (Rettberg and Thomas 1986) and the Sequent Balance system (Osterhaug 1985). Algorithms designed for multiprocessors are known as asynchronous parallel algorithms. An asynchronous algorithm is a collection of processes, some or all of which are executed simultaneously on a number of available processors. When the parallel algorithm begins execution on a processor, it creates a number of processes to be performed. If free processors are available, these processes are assigned to the processors to perform the necessary computations. Otherwise, a process is queued and waits for a processor to be free. When a processor completes execution of a process, it becomes free.

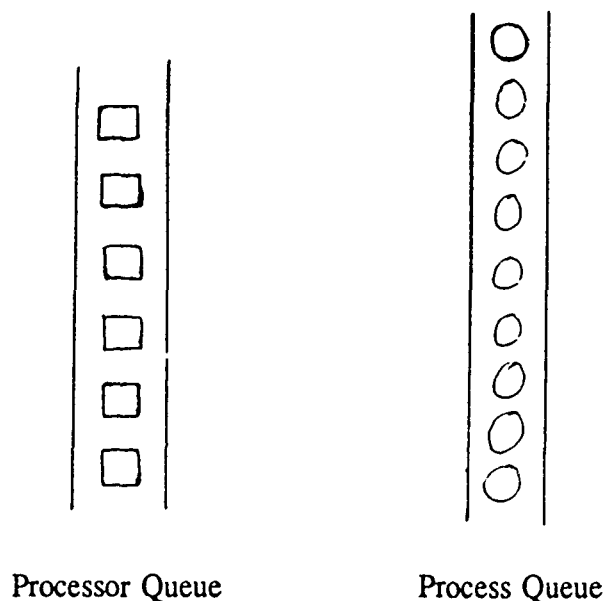


Figure 2. Two Main Queues of a Multiprocessor

Multiprocessors are very popular because they are easy to program. Multiprocessors work with two queues: a process queue and a processor queue. The main thing a programmer has to do is create an optimal number of processes out of a program. The processes are queued in the process queue. Free processors wait in the processor queue. If processes are waiting in the process queue then they are assigned to available processors from the processor queue. After a process is executed, the processor becomes free and is queued back to the processor queue.

There is another interesting but specialized parallel architecture of computation known as the neural network in which a large number of processors (similar to neurons) are interconnected and processing is done by spreading activation (McClelland et al 1986). By passing values processors change the connective strengths among themselves and propagate activations through the network. This propagation constitutes processing by the system. The neural network is an attractive model of parallel computation for building trainable systems (Howells 1988).

When fully developed GIP would be a very large and complex system that would require a hybrid architecture for efficient execution of its varied components. For example, a neural network architecture may be needed for some of the learning components of GIP; whereas a shared memory architecture could be used for the natural language processor (Dey and Hayashi 1989). Further studies would be necessary to specify the exact architecture for GIP. However, at this time, it can be concluded that GIP would require a hybrid architecture with some MIMD components, since GIP must allow asynchronous parallelism.

## 5. NATURAL LANGUAGE PROCESSING IN THE GIP ENVIRONMENT

Natural Language Processing technology is mature enough to be put in certain real world use with appropriate user assistance. Information extraction from messages is one of the most promising fields for this technology. User assistance would be required in the extraction process until some future breakthrough makes totally automated natural language understanding possible. GIP requires robust, user-oriented, modifiable, parallelizable NLP modules. GIP is recommended to have at least two alternative NLP modules so that users would be able to compare the results of the two NLP modules.

An experimental study of integrating an NLP system in the GIP environment has been conducted in the summer of 1991. The main goal of the study has been to broadly specify the requirements of the NLP modules of GIP. These requirements are briefly described below:

- (1) GIP would require well-documented NLP modules. In order to integrate the NLP modules documents like algorithms, description software design documents and user's manual would be necessary. Software development models used for the modules should be briefly described.
- (2) The NLP component should be properly modularized. The interfaces between modules should be well-defined.
- (3) The NLP modules should allow parallelization for efficient execution.

(4) Scalability should be considered as an essential feature of the NLP modules. It is often required to incorporate this provision in the design level; otherwise, the modules have to be redesigned.

(5) For messages with free text, GIP NLP modules should emphasize discourse/text analyses rather than sentential analyses. Pronominal references should be resolved in examples like:

(5.1) John has been dating Mary for the past two years. He is planning to marry her.

(6) Semantic interpretation needs to exploit different sources of Knowledge such as lexical knowledge, syntactic knowledge and domain knowledge. Prepositional phrases like "OF THE PRC AND THE SOVIET UNION" in the following fragment of text cause problems for natural language systems that do not properly integrate knowledge sources.

(6.1) LIMA, 25 OCT 89 (EFE) -- [TEXT] POLICE HAVE REPORTED THAT TERRORISTS TONIGHT BOMBED THE EMBASSIES OF THE PRC AND THE SOVIET UNION.

(7) Appropriate error handling capabilities must be provided in the modules. Whenever necessary, the control should be returned to the user.

(8) The NLP modules should be modifiable otherwise GIP will not be able to deal with the emerging technology.

(9) Most NLP systems provide a learning module at the lexical level. Some learning or adaptive capabilities must be provided in addition to the lexical level.



## BIBLIOGRAPHY

Aho, A.V. and J. D. Ullman. 1972. The theory of parsing, translation and Compiling: volume 1. Englewood Cliffs, N.J.: Prentice-Hall.

Akl, S. 1985. Parallel Sorting Algorithms, Academic Press.

Alshaw, Hiyan 1990. Resolving Quasi Logical Forms, Computational Linguistics, 16, 133-144.

Alshaw, Hiyan and J. van Eijck 1989. Logical Forms in the Core Language Engine, 27th Annual Meeting of the Association for Computational Linguistics.

Athas, W. C. and C. L. Seitz (1988) "Multicomputers: Message Passing Concurrent Computers" *IEEE Computer* 21: 9-24.

Barton, G. E., Berwick, R., and Ristad, E. 1987. *Computational Complexity and Natural Language*, MIT Press.

Beizer, B. 1990. *Software Testing Techniques*, Van Nostrand Reinhold.

Bobrow, D.G. and A. Collins 1975, (eds.) 1975. *Representation and Understanding*, New York: Academic Press.

Chomsky, N. 1957. *Syntactic structures*, Mouton, The Hague.

Chomsky, N. 1965. Aspects of the theory of syntax, Cambridge, Ma: MIT Press.

Chomsky, N. 1981. Lectures on Government and Binding, Dordrecht: Foris.

Conery, J. S., 1987, The AND/OR Process Model for Parallel Execution of Logic Programs. Kluwer Academic Publishers, Boston.

Cottrell, Garrison W. 1984, A Model of Lexical Access of Ambiguous Words, Proceedings of AAAI-84.

Dey, P., Iyengar S. S. and J. S. Byoun 1988, "Parallel Processing of Tree Adjoining Grammars," Dept. of Comput. Sci. University of Alabama at Birmingham, Report. Submitted for publication.

Dey, P., and Hayashi, Y. 1990, "A multiprocessing model of natural language processing," Theoretical Linguistics, 16, 11-23.

Dey, P., Bryant, B. and Takaoka, T. 1990, "Lexical ambiguity in tree tree adjoining grammars," Information Processing Letters, 34, 65-69.

Earley, J. 1970. "An efficient context-free parsing algorithm". Communications of the ACM 13, 94-102.

Grosz, B. 1977. The representation of use of focus in Dialogue Understanding, Ph.D. Diss. Technical Note No.151, SRI International, CA.

- Gazdar, G. 1981. "Unbounded dependencies and coordinate structure". *Linguistic Inquiry* 12, 155-184.
- Hass, A. 1990. Sentential Semantics for Propositional Attitudes", *Computational Linguistics*, 16, 1990, 213-233.
- Hillis, D. and G. L. Steele Jr. 1986. "Data Parallel Algorithms," *Communications of the ACM*, 29, 1170-1183.
- Hirst, G. J. 1984 *Semantic Interpretation against ambiguity*, Ph.D. diss. Brown University.
- Howe's, T. (1988) "Vital - A Connectionist Parser" *Proceedings of the 10th Annual Meeting of the Cognitive Science Society*, Canada.
- Jackendoff, R. S. 1972. *Semantic Interpretation in generative grammar* , MIT Press.
- Jamieson, L. H., D. B. Gannon and R. J. Douglass (eds). 1987. *The Characteristics of Parallel Algorithms*, MIT Press.
- Joshi, A. K. 1985. "Tree Adjoining Grammars: How much context-sensitivity is required to provide reasonable structural descriptions?" Dowty, et al (eds.) *Natural Language Parsing*.
- Kaplan, D. 1975. "Quantifying in", in Davidson et al eds. *The Logic of Grammar*,
- Keenan, E. L. and L. M. Faltz 1985, *Boolean Semantics for Natural Language*, Dordrecht: D. Reidel.
- Kowalski, R. A., 1974, "Predicate Logic as a Programming Language." In *Information Processing 74*, IFIP No. 5, 733-742.
- KSC 1991. *Generic Intelligence Processor (GIP)*, Technical Proposal, Volume III, Knowledge Systems Concepts Inc. , Rome, NY.
- Kuck, D. J. 1977, "A Survey of Parallel Machine Organization and Programming," *ACM Comput. Survey*, 9, 29-59.
- Kumar, V. and Kanal, L. N. 1984, "Parallel Branch and Bound Formulations for understanding and synthesising AND/OR TREE search. *IEEE Trans. Pattern Analysis and Machine Intell.* 6, 768-778.
- Kung, H. T. 1980. *The Structure of Parallel Algorithms*. Adv. Comput. 19, New York: Academic Press, 65-112.
- Lenat, D. et al 1990. "CYC: Toward Programs With Common Sense", *Communication of the ACM*, 33, 30-49.
- Marcus, M. 1980. *A theory of syntactic recognition for natural language*. Cambridge, MA: MIT Press.
- May, R. 1985. *Logical Form*. Cambridge, MA: MIT Press.

- McCawley, J. D. 1968. "The role of semantics in a grammar". In E. Bach and R.T. Harms (eds), *Universals in Linguistic Theory*, New York: Holt, Rinehart and Winston.
- Milne, R. 1986. Resolving Lexical Ambiguity in a Deterministic Parser", *Computational Linguistics*, 12, 1-12.
- Morell, L. 1990. A Theory of Fault-based Testing, *IEEE transactions on Software Engineering*, 844-857.
- Ottmann, T. A., Rosenberg, A. L. and J. L. Stockmeyer. 1982, "A Dictionary Machine (for VLSI)," *IEEE Transactions on Computers*, 31, 9, 892-897.
- Raynal, M. 1985. *Algorithms for Parallel Processing*, MIT Press.
- Schank, R. and L. Birnbaum 1980. "Memory, Meaning, and Syntax," Report 189, Dept. of Comp. Science, Yale Univ.
- Schmidt, D. A., 1986, *Denotational Semantics*, Allyn and Bacon, Inc., Boston.
- Schubert, L. and F. J. Pelletier 1982, "From English to Logic: Context-Free Computation of 'Conventional' Logical Translation," *Computational Linguistics* 8, 27-44.
- Snider, C. 1979. Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. Ph.D. Diss. MIT.
- Somani, A. K. and V. K. Agarwal. 1984, "An efficient VLSI dictionary machine," *Proceedings of the 11th Annual ACM Int. Symp. Comput. Architecture*, 142-150,
- Steinberg, D. and L. A. Jakobovits (eds) 1971, *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*.
- Stockwell, R. P., Schachter, P. and B. H. Partee 1973, *The Major Syntactic Structures of English*. New York: Holt, Rinehart and Winston.
- Tennant, H. 1981. *Natural language processing*. New York: Petrocelli.
- Voas, J., Morell, L. and K. Miller 1991. Prediction Where Faults can Hide from Testing, *IEEE Software*, 41-48.
- Waltz, D. and J. Pollack. 1984. "Phenomenologically Plausible Parsing", *AAAI-84*. 335-339.
- Warren, D. S. and J. Friedman 1982. "Using Semantics in Non-Context-Free Parsing of Montague Grammar," *Computational Linguistics* 8, 123-138.
- Warren, H. D. and F. C. N. Pereira 1982, "An Efficient Easily Adaptable System for Interpreting Natural Language Queries," *Computational Linguistics* 8, 110-122.

Webber, B. 1979. *A Formal Approach to Discourse Anaphora*, New York: Garland.

Winograd, T. 1972. *Understanding Natural Language*, New York: Academic Press.

Winograd, T. 1983. *Language as a cognitive process: Syntax*. Reading, Mass.: Addison-Wesley.

Woods, W.A. 1970. "Transition Network Grammars for Natural Language Analysis". *Communications of the ACM*, 13, 591-606.

Woods, W.A. 1968. "Procedural Semantics for a question answering machine", Fall Joint Computer Conference, 457-471.

#### ACKNOWLEDGEMENT:

This research was conducted at Rome Laboratory, Griffis AFB, Rome, NY with support from U.S. Air Force under the AFOSR 1991 program. I am grateful to Walter Gadz, John Pirog, John Salerno, Michael Thomas, Mary Ellis, Rob Zeigler and many others for their comments, help and encouragement.

## APPENDIX: Programs

/\*

~ gip/remotecompare.c

Peter Dey

This function copies interactively an entire directory from another site using rcp . It prompts for a remote pathname that has to be copied.

See also getgould.c

See also remotenlp.c

\*/

# include "stdio.h"

# include "stdlib.h"

main()

{

FILE f;

int status;

printf(" INTERACTIVE MESSAGE ACQUISITION FROM REMOTE MESSAGE-SOURCE n");

printf(" THE NOTATION FOR ENTERING MESSAGE-SOURCE IS: username@site:directory n n");

printf(" ENTER THE MESSAGE-SOURCE: ");

scanf("%s", &f);

if (fork() == 0)

execl("/usr/ucb/rcp", "rcp", "-r", f, "Demo", NULL);

wait(&status);

if (fork() == 0)

printf("Message Acquisition Complete0);

wait(&status);

}

```

/*
~gip/remotenlp.c
Peter Dey
This program sends messages to another site using rcp.
Then, it processes them on that remote site using remote procedure
calls. Finally it transmits the results back to the this site.
See also ~gip/remoteacquire.c
*/

# include "stdio.h"
# include "stdlib.h"

main() {

int status;
if (fork() == 0)
execl("/usr/ucb/rcp", "rcp", "Demo/testdir2/ms1", "gip@buckwheat:ms1", NULL);
wait(&status);
if (fork() == 0)
execl("/usr/ucb/rcp", "rcp", "Demo/testdir2/ms2", "gip@buckwheat:ms2", NULL);
wait(&status);
if (fork() == 0)
execl("/usr/ucb/rcp", "rcp", "Demo/testdir2/ms3", "gip@buckwheat:ms3", NULL);
wait(&status);
if (fork() == 0)
execl("/usr/ucb/rsh", "rsh", "buckwheat", "muc3test", "ms1", NULL);
wait(&status);
if (fork() == 0)
execl("/usr/ucb/rsh", "rsh", "buckwheat", "muc3test", "ms2", NULL);
wait(&status);
if (fork() == 0)
execl("/usr/ucb/rsh", "rsh", "buckwheat", "muc3test", "ms3", NULL);
wait(&status);
if (fork() == 0)
execl("/usr/ucb/rcp", "rcp", "gip@buckwheat:ms1.fill", "ms1.fill", NULL);
wait(&status);
if (fork() == 0)
execl("/usr/ucb/rcp", "rcp", "gip@buckwheat:ms2.fill", "ms2.fill", NULL);
wait(&status);
if (fork() == 0)
execl("/usr/ucb/rcp", "rcp", "gip@buckwheat:ms3.fill", "ms3.fill", NULL);
wait(&status);
execl("/usr/ucb/cat", "cat", "ms1.fill", "ms2.fill", "ms3.fill", NULL);

}

```

```

/*
/home/gip/demol.c
Peter Dey
This program uses suntools to make an on-line slide. The slide can be
used in the gip environment during a gip demo. Compile the file by:
cc -o demol demol.c -lsuntool -lsunwindow -lpixrect
After compilation the object-code would be in demol. Enter the
gip demo environment and then execute demol via a shell tool.
*/

#include <stdio.h>
#include <suntool/sunview.h>
#include <suntool/panel.h>
#include <suntool/icon.h>

main( )
{
    Frame frame;
    Panel panel;
    Pixfont *bold;
    bold = pf_open("/usr/lib/fonts/fixedwidthfonts/cour.b.24");
    if(bold == NULL) exit(1);
    frame = window_create(NULL, FRAME, FRAME_LABEL, "                GIP", 0);
    panel = window_create(frame, PANEL, WIN_FONT, bold, 0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, "    GENERIC INTELLIGENCE    PROCESSOR (GIP)",
                                0);
    window_fit(panel);
    window_main_loop(frame);
}

```



*P*  
*/home/gip/demo2.c*  
 Peter Dwy  
 This program uses Suntools to make a on-line slide. The slide can be  
 used in the gip environment during a gip demo. Compile the file by:  
 cc -o demo2 demo2.c -lsuntool -lsunwindow -lirect  
 After compilation the object-code would be in demo2. Enter suntools  
 or the gip demo environment and then execute demo2.  
 \*/

```
#include <stdio.h>
#include <suntool/sunview.h>
#include <suntool/panel.h>
#include <suntool/icon.h>

main()
{
    Frame frame;
    Panel panel;
    Fontset *bold;
    bold = pf_open("/usr/lib/fonts/fixwidth/fonts/cour.b.24");
    if(bold == NULL) exit(1);
    frame = window_create(NULL, FRAME, FRAME_LABEL, "GIP", 0);
    panel = window_create(frame, PANEL, WIN_FONT, bold, 0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, "  WHAT IS THE GENERIC INTELLIGENCE PROCESSOR (GIP) ? "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, "=====> "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " ==> A TOOLKIT FOR DEVELOPING, TESTING AND DELIVERING "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " MESSAGE PROCESSING APPLICATIONS "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " IT CONSISTS OF: "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, "=====> "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " ", 0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " ==> USER SELECTABLE INPUT SOURCES "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " ==> AUTOMATIC/USER ASSISTED INFORMATION EXTRACTION "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " ==> AUTOMATIC OUTPUT FORMATTING FOR DEVELOPMENT SYSTEMS "
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " ==> UTILITIES FOR DEVELOPING/TESTING NEW PROCESSING ALGORITHMS"
        0);
    panel_create_item(panel, PANEL_MESSAGE,
        PANEL_LABEL_STRING, " ", 0);
    window_fit(panel);
    window_main_loop(frame);
}
```

```

/*  /usr/etc/helpgip.c
    Peter Dey
    This program prints out hints for running gip
*/

#include "stdio.h"
#include "stdlib.h"

main()
{

printf("1. Login: gip \n");
printf("2. Wait & Click as directed \n");
printf("    (Execute demol from shell tools, optionally) \n");
printf("    (Execute remoteacquire on CONSOLE, optionally) \n");
printf("3. Click on PROJECT - Open & Accept \n");
printf("4. Click on DEVELOPER - Open & Accept \n");
printf("5. Click on DEVELOPER - Layout \n");
printf("6. Click on ACQUIRE - (Change Fact Base Name, optionally) \n");
printf("7. Click on CONTROL - Run \n");
printf("    (Execute remotenlp on CONSOLE, optionally) \n");
printf("8. TO QUIT: Click on LADIES - Quit \n");
printf("    Click on Gray - exit \n");

}

```

**1991 USAF-RDL SUMMER RESEARCH PROGRAM**

**MILLIMETER-WAVE NOISE  
MODELING INVESTIGATION**

**Sponsored by the  
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH  
Administered by the  
Research and Development Laboratories**

**FINAL REPORT**

Faculty participant:	Lawrence P. Dunleavy, Asst. Professor
Graduate student participant:	Steven M. Lardizabal
University Affiliation:	Department of Electrical Engineering University of South Florida Tampa, FL 33620
USAF Researcher:	Richard Webster Rome Laboratories, RL/EEAC Hanscom AFB, MA 01731
Date:	July 19, 1991
Contract No.:	F49620-90-C-0076

# **MILLIMETER-WAVE NOISE MODELING INVESTIGATION**

## **ACKNOWLEDGEMENTS**

This work was sponsored under the United States Air Force AFOSR Summer Research Program, (Contract #F49620-90-C-0076) administered by Research and Development Laboratories.

The encouragement and assistance of Mr. R. Webster and Dr. P. Carr, and the endorsement and support of Dr. J. Schindler and Dr. R. Mailloux, and Lt. Col. R. Momberg are gratefully acknowledged.

We would also like to acknowledge the helpful insights gained through conversations with the following noise modeling experts: Dr. R. Pucel of Raytheon Company, Dr. M. Pospieszalski of the National Radio Astronomy Observatory, Dr. M. Gupta of Hughes Aircraft Company, and Mr. V. Adamian of Automatic Testing and Networking (ATN) Company. ATN also helped provide the noise measurements reported here.

# **MILLIMETER-WAVE NOISE MODELING INVESTIGATION**

## **ABSTRACT**

A new method is described that provides valid mm-wave noise models for field effect transistors, including MESFETs and HEMTs. This method avoids the complications of a variable impedance tuner, and requires only the knowledge of a small signal equivalent circuit, and noise figures measured across a range of frequencies for a single known source impedance. Noise parameters derived from this method are shown to agree well with those obtained from tuner based measurements. A review of previously reported noise modeling techniques, summarized here, reveals either their common dependence on tuner based measurements or the use of approximations that are not valid at mm-wave frequencies.

# MILLIMETER-WAVE NOISE MODELING INVESTIGATION

## I. INTRODUCTION

It is well known that the use of a low noise amplifier (LNA) at the front end of a mm-wave radar or communications receiver can improve system sensitivity and dynamic range as compared to the use of down conversion with a mixer immediately following the antenna. LNAs are now possible at frequencies as high as W-band (75-100GHz), using state-of-the-art High Electron Mobility Transistors (HEMTs). However, accurate transistor noise characterization is critical to achieving good millimeter-wave (mm-wave) LNA performance, while avoiding the widely used process of hand tuning. Such hand tuning for noise performance, costly and time consuming for hybrid amplifiers, is simply not practical for millimeter-wave monolithic microwave integrated circuits (MMICs).

As illustrated in Figure 1, FET noise characterization can take the form of measured noise parameters ("Black Box Methods") or a noise model ("FET Specific Methods"), from which noise parameters can be calculated. A knowledge of four noise parameters allows for a calculation of the device's noise figure for a given complex input terminating admittance, according to [7]:

$$F = F_{\min} + \frac{R_N}{G_s} |Y_s - Y_{\text{opt}}|^2 \quad (1)$$

where  $F$  is the noise figure of the two-port device,  $Y_s$  is the input or source admittance presented to the device, and  $G_{\text{opt}}$ ,  $B_{\text{opt}}$ ,  $F_{\min}$ , and  $R_N$ , are the four noise parameters defined as follows:

$Y_{opt} = G_{opt} + jB_{opt}$  is the optimum input admittance<sup>1</sup>

$F_{min}$  is the minimum noise figure achieved for  $Y_s = Y_{opt}$

$R_N$  is called the equivalent noise resistance, and describes how rapidly the noise figure degrades as the source admittance moves away from  $Y_{opt}$ .

Noise parameter measurements are relatively routine at microwave frequencies, using conventional techniques [1-4]. However, progress has been slow towards the implementation of mm-wave noise parameter measurements. This is mainly due to the difficulty with the fabrication and characterization of variable impedance tuners at mm-wave frequencies as discussed in Section II.

The present research has explored FET specific noise model methods (Figure 1) with the goal of providing an alternative or complementary mm-wave FET noise characterization method to those relying on tuner based measurements. Accordingly, an overview of FET noise modeling methods is given in Section III. Also described (Section IV) is a new method developed for deriving valid mm-wave FET noise models through the use of an equivalent circuit and a set of noise figure measurements over frequency. Figure 2 shows the model topology used.

## II. OBJECTIVES AND ACCOMPLISHMENTS

At the start of the summer's work, a short "Objectives and Summary Report containing the following information was submitted to our Air Force sponsors. This provided the template for the Summer's work.

**PROBLEM:** Conventional microwave noise modeling methods utilize a variable impedance (or admittance) tuner, which presents a sequence of "known" source impedances  $Z_S$  (admittances  $Y_S$ ) to a device's input terminals. The tuner must be able to present a range of impedances that represent a good coverage of the Smith Chart, including near the edges of the chart. Hence, the tuner must have very low loss. The ability to fabricate, and characterize, such a tuner becomes increasingly more difficult as frequency is increased. Research and commercial developments are being pursued in the industry and at the Air Force<sup>2</sup> implementing mm-wave tuners for noise parameter measurements. Because of the inherent difficulties with this approach, however, it is essential that alternatives be explored to the reliance on a variable impedance tuner for mm-wave FET noise characterization.

**SCOPE OF RESEARCH:** This research consisted of a ten week effort involving the work of two researchers. The work involved a paper study, analytical research, and numerical experiments. The numerical experiments combine analytical research with existing measurement data in an attempt to explore the validity of a new noise modeling approach (see Section IV ). Some of the numerical experiments were carried out by modifying existing computer programs generated by a previous Summer Research Associate. A new computer program was also generated as part of this work. Because of the time constraints, no attempt was made to perform measurements for use with this study, instead existing data was made available to us from the industry.

**SPECIFIC OBJECTIVES VS. ACCOMPLISHMENTS:** The work described above was carried out as planned and all of the objectives of the



work were satisfied. This degree of accomplishment was aided by extensive interaction with, Mr. Rick Webster. This included short weekly reports and weekly meetings.

The research was divided into several tasks and accomplished as follows:

1) OBJECTIVE - Review and assess the previous MESFET noise modeling work of Mr. Bill Patience [5], as performed under the 1989 AF Summer Research Program.

ACCOMPLISHMENT - After verifying the theoretical derivations of this work, an assessment was made with regard to the theoretical assumptions made, including that of frequency independence for noise source coefficients (P,R, and C) used with the model of Figure 2. This assessment was made from theoretical considerations, and numerical case studies derived from measurement data (data which was not available to Mr. Patience at the time of his study). See Section IV<sup>3</sup>.

2) OBJECTIVE - Explore improvements and alternatives to the use of tuner based measurements for noise characterization.

ACCOMPLISHMENT - A paper study was conducted to formulate an overview of various noise modeling methods documented in the literature (Section III). This study proposes a modification to the work of Patience as the most promising mm-wave noise modeling method, that does not require tuner based measurements. A related new computer program was generated.

3) OBJECTIVE - Prepare a final report in accordance with AFOSR Summer Research program guidelines, and a "mini-grant" proposal for a follow-on research project.

ACCOMPLISHMENT - The present document satisfies the final report requirements, and a related oral presentation was given at Hanscom AFB July 12, 1991. Because this work is of interest to other Air Force Laboratories, presentations of the work are scheduled to be given at Griffiss AFB July 22, 1991, and at Wright Patterson AFB August 15, 1991. The completion of the mini-grant proposal is forthcoming.

4) OBJECTIVE - Prepare a draft (or summary) version of a research paper aimed at publication in a trade journal (e.g. *IEEE Trans. on Microwave Theory and Tech.*), a trade magazine (e.g. *Microwave Journal*), or a conference digest (e.g. *IEEE MTT-S Symposium Digest*).

ACCOMPLISHMENT - An abstract for a paper entitled "A New Millimeter-wave Noise Modeling Method for Field Effect Transistors," was submitted to the 1991 IEEE International Electron Devices Meeting (IEDM). If accepted, an expanded four page paper will be required for the conference digest. It is also expected that a second paper will be generated that gives a comprehensive review of FET noise modeling methods.

### III. A BRIEF OVERVIEW OF FET NOISE MODELING METHODS

A viable approach for mm-wave noise characterization of FETs is to use knowledge of an equivalent circuit for the FET to reduce the required noise measurement complexity. Several such FET specific methods have been

proposed in the literature (Figure 1). A brief summary of these methods is given here.

The reported FET noise modeling methods [7-30] use various approximations, but follow from a common theoretical formulation. This theoretical formulation is based upon the fundamental linear two port noise theory developed by Rothe and Dahlke [6] and further developed by an IRE Subcommittee on Noise [7]. This two-port noise theory applies to both Black Box noise parameter measurements and FET Specific Noise Models as indicated in Figure 1.

**EARLY FET NOISE THEORY:** The early groundwork for FET noise modeling was laid by van der Ziel [8-10]. van der Ziel [13], identified the main source of noise in FETs as thermal noise in the channel. He derived useful analytical expressions for the intrinsic gate and drain current sources ( $i_{ng}$  and  $i_{nd}$ ) and the correlation coefficient ( $C$ ) between them. Under his approximations,  $C$  was found to be imaginary. Klaassen, in a slightly more general formulation [11], found  $C$  to be complex in general with a small real part, which he attributed to high frequency gate-channel coupling (neglected by van der Ziel). Nonetheless van der Ziel's work, complete with imaginary  $C$ , has provided the format for nearly every noise modeling solution to follow.

**IMPORTANT NOISE THEORY EXTENSIONS:** Important extensions to this early work were made by Baechtold [12-13], Baechtold advanced a complete noise modeling procedure, that relies on frequency independent noise coefficients  $P$ ,  $R$ , and  $C$  that follow directly from van der Ziel's relations [8-

9]. The intrinsic noise sources are expressed in terms of P,R, and C through the following relations:

$$\begin{aligned} \overline{i_{nd}^2} &= 4kT\Delta f g_m P & \overline{i_{ng}^2} &= \frac{4kT\Delta f C_{gs}^2 \omega^2}{g_m} R & jC &= \frac{\overline{i_g i_d}}{\sqrt{\overline{i_{nd}^2} \overline{i_{ng}^2}}} \end{aligned} \quad (2)$$

This same noise source description is used in the present research (Figure 2). In Baechtold's method, noise parameters are calculated using a small signal model, fitted to measured S-parameters combined with analytically determined noise coefficients P,R and C.

Another very important contribution was made by Pucel et. al. [14]. This work covers all aspects of MESFET operation, including DC, small signal, and noise characteristics. Pucel et. al. include a review is given of prior work in each of these areas and point out that the previous FET noise theory (including that of van der Ziel and Baechtold) fails to adequately account for velocity saturation effects. The authors proceed to use a two section velocity field model, valid in the saturation region to derive expressions for P,R, and C and for the four noise parameters. This method, like that of Baechtold's, is not directly applicable for accurate millimeter-wave noise characterization due to the omission of parasitic elements  $C_{dg}$  and  $R_{ds}$  (see Figure 2) from the analysis. In addition, this model shares the limitations of other purely analytic/numerical models discussed below.

The remaining FET noise model work (Figure 1) has been placed into three categories: 1) Analytical/Numerical Noise Models, 2) Semi-empirical Noise Models, and 3) Physically Based Theoretical/Experimental Models.

***ANALYTICAL /NUMERICAL NOISE MODELS:*** The first category (Figure 1) includes the work of van der Ziel, Baechtold, and Pucel et. al., as well as the subsequent work of Cappy et. al. [15-16], Heinrich [17-18], and Brookes [19]. These methods apply basic physics to calculate noise properties, using as inputs the geometry and DC operating conditions of the FET. Such models are very useful for establishing physically meaningful model topologies and studying trends in electrical behavior with changes in geometry and operating conditions. However, analytic/numerical methods, by themselves, do not produce the most accurate FET models for circuit design. This is due to the many non-ideal conditions existing in practical FETs including, for example, variations in the geometrical dimensions and doping levels. This information is also seldom available to the circuit designer. Hence, a practically useful FET noise model requires that at least some of the noise model parameters be determined through the use of measured data.

***SEMI-EMPIRICAL NOISE MODELS:*** The second category of noise models (Figure 2) uses measured data to determine empirical noise fitting factors. These fitting factors are used along with a simplified equivalent circuit to estimate noise parameters. Most notable of these are the models proposed by Fukui [20-22], and Podell[23]. These models are useful in that they provide convenient closed form relations between dominant geometrical or equivalent circuit parameters and noise performance, yet because of the many

approximations made they are not well suited for mm-wave FET noise characterization.

### **PHYSICALLY BASED THEORETICAL/EXPERIMENTAL MODELS:**

The third category (Figure 1) contains the most promising methods for mm-wave applications. This work includes the methods of Gupta et. al. [24-25], Pospieszalski et. al. [26-27], Robertson et. al. [28-30], and Riddle [31]. All of these methods assume (or equivalently assume) that the coefficients  $P$ ,  $R$ , and  $C$  are independent of frequency and rely on the use of measured  $S$ -parameters to derive a small signal equivalent circuit. The methods differ in the characterization of the noise sources within the model, in the simplifying assumptions and the exact model topologies.

Gupta et. al. [24-25] use a simplified model that only requires the output noise power spectral density that is determined through an output noise power measurement made at a single low frequency (e.g. 1-2GHz). Pospieszalski et. al. [26] use a more general model that requires a set of noise parameters at a single frequency to characterize equivalent temperatures  $T_g$  and  $T_d$  used to represent the gate and drain noise sources respectively. Robertson et. al. [28-30] also use a fairly general model, similar to that of this research (Figure 2). They determine the noise coefficients  $P$ ,  $R$ , and  $C$  by fitting an equation for  $F_{min}$ , given in terms of  $P$ ,  $R$ , and  $C$  to a set of measured  $F_{min}$  values determined for a number of frequencies ( $\geq 4$ ). Most recently, Riddle [31] presented a means to directly extract  $P$ ,  $R$ , and  $C$  values for a model similar to Figure 1 by using sequential matrix manipulations of the measured single frequency noise parameters and  $S$ -parameters.

Of these "Physically Based Theoretical/Experimental Noise Models," methods, the most extensively implemented and verified is that of Pospieszalski's [27], and this method has been applied to develop noise models for use at cryogenic temperatures [26]. Nonetheless, while each of the methods discussed have merit, due to various approximations they are generally not valid at millimeter-wave frequencies, or if valid require tuner based measurements in order to characterize the internal noise sources of the model.

#### **IV. A NEW MILLIMETER-WAVE NOISE MODELING METHOD**

The method advanced as part of the present research, overcomes the difficulties of the above methods. The new method makes use of noise figure data measured over a range of frequencies along with the equivalent circuit parameters to determine three frequency independent noise coefficients (P,R, and C). This technique avoids unnecessary simplifications to the equivalent circuit model and remains valid at millimeter-wave frequencies.

**METHOD OF PATIENCE:** The new method builds on the method described by W. Patience [5], which is illustrated in the algorithm of Figure 3. The method requires a small signal model, derived from S-parameter measurements, and noise figure data taken over a range of frequencies for a single known source reflection coefficient condition  $\Gamma_s$ , also measured over frequency. Central to the algorithm is an equation for the noise figure expressible in the following form:

$$F = M_1 + M_2 R + M_3 P + M_4 \sqrt{RP} \quad (3)$$

where  $M_i$ ,  $i = 1, 2, 3, 4$ , are functions of the known equivalent circuit parameters, and measured values of  $\Gamma_s$ , corresponding to each measured value of noise

figure supplied. The unknowns,  $P$ ,  $R$ , and  $C$ , are determined using a least squares fitting algorithm [2], equation (3), and the measured data. Once they have been determined, a synthesized tuner based "measurement" procedure computes the noise parameters at any desired frequency. The noise figure and reflection coefficient data required by this algorithm can readily be obtained from on-wafer or fixtured measurements even at millimeter-wave frequencies. An example of an applicable mm-wave measurement system and procedure is described by Dunleavy [32].

**EXAMINATION OF THE ALGORITHM:** The present research included a thorough investigation of the method of Figure 3, through both theoretical verification and numerical experiments that utilize measured data to investigate and verify the method. From a theoretical point of view, the derivation of Patience was found to be sound, except for one assumption that the term  $Z_{cor}$  (see Roth and Dahlke [6]) was assumed to be equal to the network parameter  $Z_{11}$ . This assumption is believed to be an unnecessary constraint.

Next, the Patience method was investigated on the basis of a set of numerical experiments made using measured data provided from industry for a MESFET. The measured data, the output from a commercially available noise parameter measurement system, consisted of measured noise figures with corresponding values of  $\Gamma_s$  at 12 frequencies, for 16 different source states per frequency. The algorithm (Figure 3) was then applied in two different ways to generate the data shown in Figure 4.



Figure 4a shows the results by applying the method using, for each source state, the noise figure data and  $\Gamma_s$  values provided versus frequency. Note that for perfect data all of these curves would be flat. The fact that they are not is due in part to errors in the measurement data. For example, the data corresponding to source state 2, 9, and 16 of Figure 4a appears questionable since C becomes greater than unity<sup>4</sup>. This illustrates how the method may be used to complement a noise parameter measurement procedure, by identifying problem source states. Some deviations may also be due to errors in the small signal model.

In Figure 4b are shown the results of applying a slightly modified algorithm at each frequency that uses the measured noise figures and  $\Gamma_s$  values provided as a function of source state. With accurate data, the experiment of Figure 4b can be used to investigate the frequency independence assumption for P, R, and C. Although most reported noise modeling methods have required or assumed P, R, and C to be frequency independent. Some numerical work has suggested that they may be frequency dependent in general [15-18]. This is a subject of continuing controversy<sup>5,6</sup>, and more data needs to be examined before reaching any definitive conclusions.

Despite the variations in the P, R, and C values observed in Figure 4, the results for noise parameters determined from the algorithm of Figure 3 agree very well with measured noise parameter measurements as shown in Figure 5. This shows considerable promise for the method's application to develop mm-wave noise models with the improvements discussed next.

**METHOD IMPROVEMENTS AND EXTENSIONS:** Based on this research several improvements to the method have been initiated. These include the derivation of a more general equation for the noise figure that does not assume  $Z_{cor} = Z_{11}$ . The results of this derivation have now been implemented into the new method and verified with related software<sup>7</sup>. Other improvements initiated are the replacement of the synthesized "measurement" step from the algorithm of Figure 3 with a direct calculation of noise parameters, and the ability to solve for the noise source coefficients in different ways depending on the available measurement data. These ways include the use of multiple source state data to calculate P, R, and C (as implemented in Figure 4b), the use of  $F_{min}$  data versus frequency (similar to Robertson et. al. [28-29]), and the use of measured noise parameters at a single frequency or at multiple frequencies. This last set of data is the output of a commercial noise parameter measurement system. It follows that a physically based model, such as the present one, fitted to the entire set of available data may provide the most accurate frequency extrapolatable noise model.

Other issues to be considered in future research are the implementation of an improved small signal model extraction algorithm, and the application of the method to temperature dependent FET noise model development. The derivation of a small signal model, as done for the present work, from a single set of S-parameters at the operating bias, may not adequately separate the extrinsic and intrinsic model parameters. To address this problem, several methods have been proposed to determine a more accurate, physically meaningful, small signal FET model [33-34]. The University of South Florida is also investigating this problem as part of a separate project. In future

developments of this noise modeling method an improved small signal model procedure should be employed. Another area of strong interest to both the industry and the Air Force is the development of temperature dependent noise models for use in CAD and for use in predicting and explaining device noise behavior at temperatures from cryogenic (e.g. as low as 4K) to elevated temperatures (e.g. 475K). Consideration should, therefore, also be given to temperature dependent noise model development in future research.

## **V. SUMMARY AND CONCLUSIONS**

This research project has been very successful in furthering the understanding of mm-wave noise modeling methodology. All of the research goals have been satisfied.

After a detailed literature review, it is concluded that the most viable methods for mm-wave FET noise modeling use a physically based noise equivalent circuit. Conventional noise parameter measurements are complicated at mm-wave frequencies, by the difficulty of fabrication and characterization of a low loss variable impedance tuner. This summer's research has advanced a new method that overcomes this difficulty by using a physically based noise equivalent circuit whose internal noise sources are characterized by noise figure measurements, made over a range of frequencies without the need for a tuner. The new method, therefore, may be used as a stand alone mm-wave noise modeling tool. Alternatively, the method may be used to complement a tuner based noise parameter measurement procedure in the following two ways. First, it can be used as a means to examine the integrity of the noise data corresponding to a particular source impedance tuner state. Second, it can be

used as a means to use the available measured noise data from the tuner procedure to derive a convenient frequency extrapolatable model.

This method builds upon the work of a previous AFOSR summer research project. An investigation of this previous work produced promising comparisons to measured data, but also revealed that modifications are necessary for reliable mm-wave model development. A new computer program that incorporates some of these improvements has been developed and verified. Work towards implementing the remaining improvements has been initiated. It is strongly encouraged that additional research be performed to more fully develop and verify the method advanced here, and to broaden its range of applicability to include temperature dependent FET noise characterization.

## REFERENCES

- [1] H. Haus ed., "IRE Standards on Methods of Measuring Noise in Linear Twoports, 1959," *Proc. IRE*, vol. 48, pp60-68, Jan. 1960.
- [2] R. Lane, "The Determination of Device Noise Parameters," *Proc. IEEE*, vol. 57, pp1461-1462, Aug. 1969.
- [3] V. Adamian, A. Uhler Jr., "A Novel Procedure for Receiver Noise Characterization," *IEEE Trans. on Instrumentation and Meas.*, pp 181-182, June 1973.
- [4] M. Sannino, "On the Determination of Device Noise and Gain Parameters," *Proc. IEEE*, vol. 67, pp1364-1366, Sept. 1979.
- [5] W. Patience, "A Simplified Method of Determining Noise Parameters of High Frequency MESFET's," 1989 USAF-AFOSR Summer Faculty Research Program Final Report, Sept. 6, 1989.
- [6] H. Rothe and W. Dahlke, "Theory of Noisy Fourpoles," *Proc. IRE*, vol. 44, pp811-818, June 1956.
- [7] H. Haus et. al., IRE Subcommittee 7.9 on Noise, "Representation of Noise in Linear Twoports," *Proc. IRE*, vol. 48, pp69-74, 1960.
- [8] A. van der Ziel, "Thermal Noise in Field-Effect Transistors," *Proc. IRE*, pp1808-1812, Aug. 1962.
- [9] A. van der Ziel, "Gate Noise in Field Effect Transistors at Moderately High Frequencies," *Proc. IRE*, pp461-467, March 1963.
- [10] A. van der Ziel and J. Ero, "Small-Signal, High-Frequency Theory of Field-Effect Transistors," *IEEE Trans. Electron Devices*, vol. ED-11, pp128-135, 1964.
- [11] F. Klaassen, "High-Frequency Noise of the Junction Field Effect Transistor," *IEEE Trans. Electron Devices*, vol. ED-14, pp368-373, July 1967.

- [12] W. Baechtold, "Noise Behavior of Schottky Barrier Gate Field-Effect Transistors at Microwave Frequencies," *IEEE Trans. Electron Devices*, vol. ED-18, pp97-104, Feb. 1971.
- [13] W. Baechtold, "Noise Behavior of GaAs Field-Effect Transistors with Short Gate Lengths," *IEEE Trans. Electron Devices*, vol. ED-19, No. 5, pp674-680, May 1972.
- [14] R. Pucel, H. Haus, and H. Statz, "Signal and Noise Properties of GaAs Microwave FET," in *Advances in Electronics and Electron Physics*, vol. 38, L. Morton, Ed. New York: Academic Press, 1975.
- [15] B. Camez, A. Cappy, R. Fauquembergue, E. Constant, and G. Salmer, "Noise Modeling in Submicrometer-Gate FET's," *IEEE Trans. Electron Devices*, vol. ED-23, pp784-789, July 1981.
- [16] A. Cappy and W. Heinrich, "High-Frequency FET Noise Performance: A New Approach," *IEEE Trans. Electron Devices*, vol. ED-36, pp403-409, Feb. 1989.
- [17] W. Heinrich, "High-Frequency MESFET Noise Modeling Including Distributed Effects," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-37, pp836-842, May 1989.
- [18] W. Heinrich, "Corrections to 'High-Frequency MESFET Noise Modeling Including Distributed Effects,'" *IEEE Trans. Microwave Theory Tech.*, vol. MTT-38, pp96-97, Jan. 1990.
- [19] T.M. Brookes, "The Noise Properties of High Electron Mobility Transistors," *IEEE Trans. Electron Devices*, vol. ED-33, pp52-57, Jan. 1986.
- [20] H. Fukui, "Optimal Noise Figure of Microwave GaAs MESFETs," *IEEE Trans. Electron Devices*, vol. ED-26, pp1032-1037, July 1979.
- [21] H. Fukui, "Design of Microwave GaAs MESFET's for Broad-Band Low-Noise Amplifiers," *IEEE Trans. Microwave Theory and Tech.*, vol. MTT-27, pp643-650, July 1979.
- [22] H. Fukui, "Addendum to 'Design of Microwave GaAs MESFET's for Broad-Band Low-Noise Amplifiers,'" *IEEE Trans. Microwave Theory and Tech.*, vol. MTT-29, Oct. 1981.
- [23] A. Podell "A Functional GaAs FET Noise Model," *IEEE Trans. Electron Devices*, vol. ED-28, pp511-517, May 1981.
- [24] M.S. Gupta, O. Pitzalis, S. Rosenbaum, and P. Greiling, "Microwave Noise Characterization of GaAs MESFET's: Evaluation by On-Wafer Low-Frequency Output Noise Current Measurement," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-35, pp1208-1217, Dec. 1987.
- [25] M.S. Gupta, and P.T. Greiling, "Microwave Noise Characterization of GaAs MESFET's: Determination of Extrinsic Noise Parameters," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-36, pp745-751, April 1988.
- [26] M. Pospieszalski, "Modeling of Noise Parameters of MESFET's and MODFET's and Their Frequency and Temperature Dependence," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-37, pp1340-1350, Sept. 1989.
- [27] M. Pospieszalski, "FET Noise Model and On-Wafer Measurement of Noise Parameters," *1991 IEEE International MTT-S Symposium Dig.*, pp1117-1120, June 1991.
- [28] R. Robertson and T. Ha, "Noise Models for Gallium Arsenide Field-Effect Transistors at Room and Cryogenic Temperatures," *Int. J. Electronics*, vol. 61, No. 4, pp443-440, 1986.
- [29] R. Robertson and T. Ha, "Optimum Noise Source Impedance Determination for GaAs FETs at Room and Cryogenic Temperatures" *Int. J. Electronics*, vol. 63, No. 3, pp359-369, 1987.
- [30] R. Robertson and R. Sanders, "A Simplified Approach to Optimum Noise Source Impedance Determination for GaAs FET Amplifiers" *Int. J. Electronics*, vol. 65, No. 5, pp943-952, 1988.
- [31] A. Riddle "Extraction of FET Model Noise-Parameters from Measurement," *1991 IEEE International MTT-S Symposium Dig.*, pp1113-1116, June 1991.
- [32] L.P. Dunleavy, "A Ka-Band On-wafer S-parameter and Noise Figure Measurement System", *34th ARFTG Conference Digest*, December 1989.
- [33] G. Dambrine, A. Cappy, F. Heliodore, and E. Playez, "A New Method for Determining the FET Small-Signal Equivalent Circuit," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-36, No. 7, July 1988.
- [34] M. Berroth and R. Bosch, "Broad-Band Determination of the FET Small-Signal Equivalent Circuit," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-38, No. 7, July 1990.

## FOOTNOTES

---

<sup>1</sup>Equation (1) is also expressible in terms of source impedance or in terms of source reflection coefficient, in which case the optimum source impedance and reflection coefficient, respectively, are the parameters of interest.

<sup>2</sup> Personal communication with R. Webster, Rome Laboratory, April 8, 1991.

<sup>3</sup>See also "R&D Record Notebooks" of L. Dunleavy, and S. Lardizabal. A separate report is also being prepared to detail the results of the numerical experiments.

<sup>4</sup> The noise sources, on a physical basis, cannot be more than 100% correlated.

<sup>5</sup> Personal communication with R. Pucel, A. Riddle, W. Heinrich at 1991 IEEE MTT-S conference.

<sup>6</sup>On another controversial point, Riddle [31] claims that  $1/f$  noise produced a significant influence on his microwave noise parameter calculations. In subsequent conversations with L. Dunleavy, both R. Pucel and M. Pospieszalski find this surprising and somewhat questionable, as  $1/f$  noise is widely believed to be entirely negligible at microwave frequencies. We may wish to re-address this question in future research.

<sup>7</sup>Results and comparisons are discussed in the separate report mentioned above detailing the numerical experiments.

# EVOLUTION OF FET NOISE CHARACTERIZATION SOLUTIONS

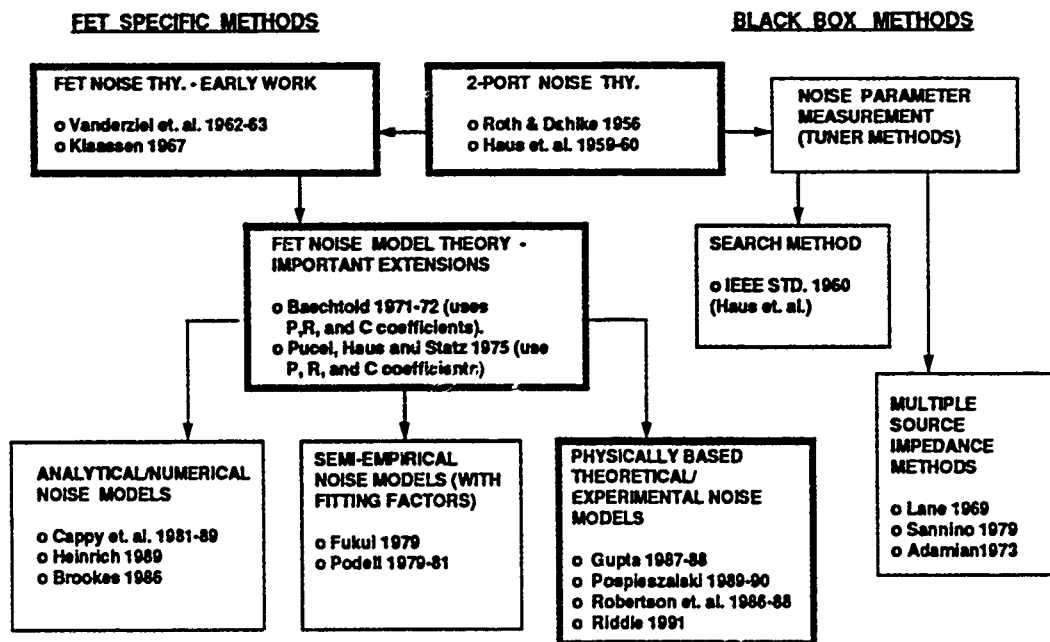
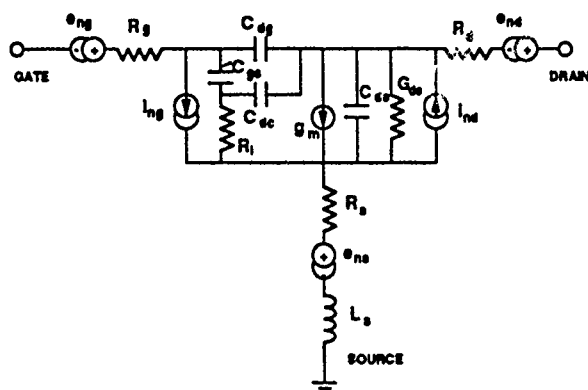


Figure 1. Flow chart showing the relationship between various reported noise characterization methods. All of the methods rely on the linear active two port noise theory developed by Roth and Dahlke.



Noise Source Description		
$\overline{e_{ng}^2} = 4kT R_g$	$\overline{e_{nd}^2} = 4kT \Delta R_d$	$\overline{e_{ns}^2} = 4kT \Delta R_s$
$\overline{i_{ng}^2} = 4kT \Delta g_{mP}$	$\overline{i_{nd}^2} = \frac{4kT \Delta C_{gs}^2 \omega^2}{g_m} R$	$Y_C = \frac{\overline{i_{ns}}}{\sqrt{\overline{i_{nd}^2} \overline{i_{ng}^2}}}$

Figure 2. Noise equivalent circuit used in the present research. Like similar circuits used by others, this circuit adds equivalent noise sources to a small signal FET model.

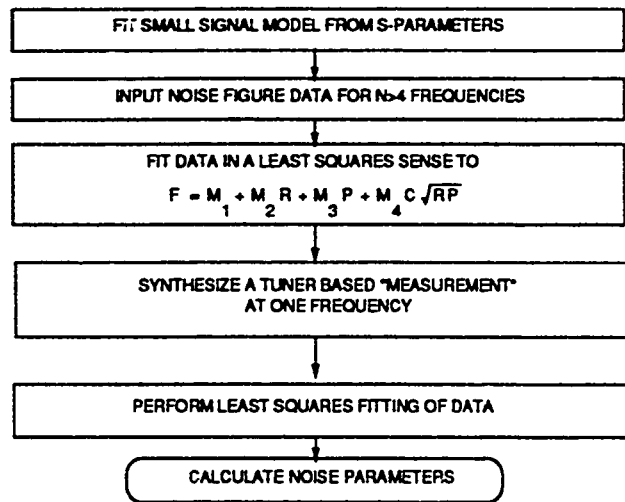
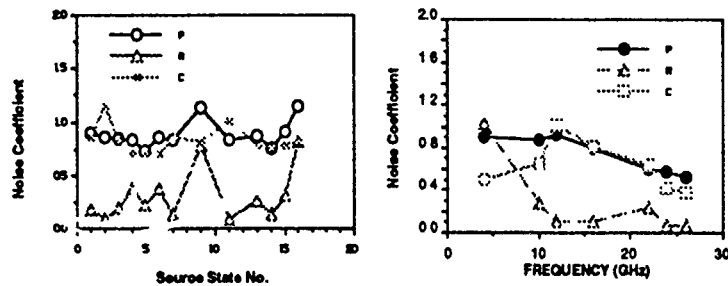


Figure 3. Algorithm for Noise Model Determination and Noise Parameter Calculation.



a. P, R, C Calculated as a function of source state.

b. P, R, C Calculated as a function of frequency

FIGURE 4. Results of numerical experiments performed with Patience's algorithm.

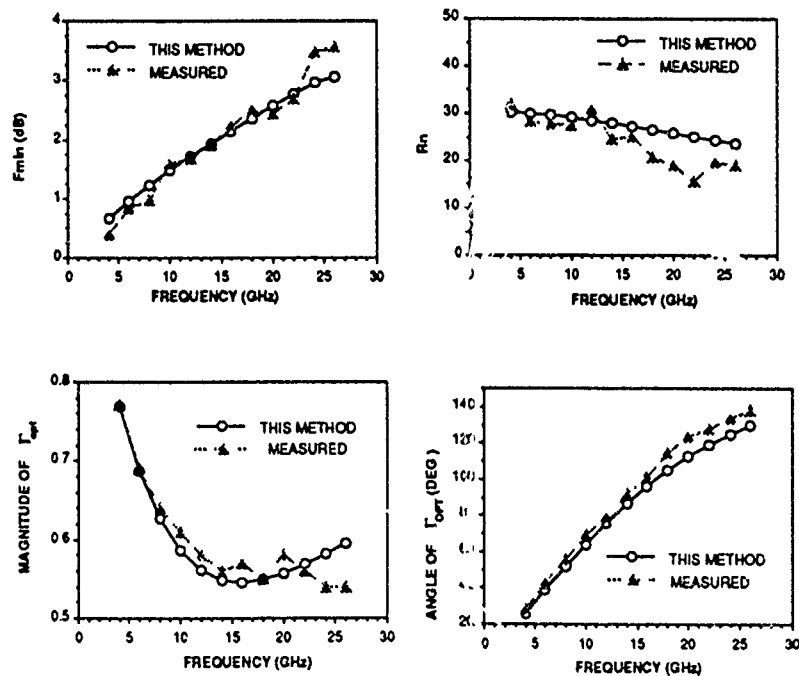


Figure 5. Comparison of noise parameters calculated for a MESFET using the present modeling method to measured noise parameters obtained with a tuner based procedure.



# **SINUSOIDAL TRANSFORM CODER PARAMETER MANIPULATION TECHNIQUES AND THEIR USE IN NETWORK AND DATA STORAGE APPLICATIONS**

*Dr. Joseph B. Evans*

Department of Electrical & Computer Engineering  
Telecommunications & Information Sciences Laboratory  
University of Kansas

## **Abstract**

This research involved increasing the capabilities of the Sinusoidal Transform Coder (STC), a high quality, low bit rate speech coder. The new methods developed in the course of this work provide a means for the compression of digital speech data for applications such as voice mail. Previous methods suitable for such applications allow only fixed compression ratios; that is, given a specified amount of input data, the stored file size must be a certain (smaller) fixed size. The speech quality is also fixed when the traditional compression methods are used. The new algorithms allow quality to be balanced against memory requirements, so that substantially greater compression can be attained, albeit at lower quality. In the new method, parameter space transformations are performed, so that the speech quality is high for a given compression ratio, as compared to alternate methods. The compression can be performed in multiple stages of any selected size, extending the previous parameter transformation techniques. Other results of this research are algorithms which provide a method for the reconstruction of lost packets of digital voice data for various loss environments were developed. The new algorithms allow very high packet loss rates to be endured, although some slight reduction in quality as compared to the equivalent unimpaired speech might be experienced. These techniques can be applied to

packet or land mobile radio environments, or a combination of both. They can also be used in conjunction with traditional error control coding methods to provide an exceptionally high degree of robustness in the presence of network and channel errors.

## **Introduction**

This report describes research performed during the AFOSR Summer Research Program related to low bit rate speech coding using sinusoidal representations. This work involved modifications and extensions to the Sinusoidal Transform Coder (STC) developed by R. J. McAulay at MIT Lincoln Laboratory under the support of the Rome Laboratories Digital Speech Laboratory [6, 8, 9, 10]. The STC modifications and extensions provide significant increases in flexibility and capability, particularly in harsh communications environments where congestion and errors are common.

This work was primarily focussed in two separate, albeit somewhat related, areas. The first of these areas is the multi-stage transformation of encoded speech from one rate to another lower rate. The second area is the reconstruction of missing packets of speech data. These areas of research are related by the method of solution, that is, manipulation of the parameteric (as opposed to waveform) representation of the speech signal.

## **Discussion of Research Problems**

Voice communications systems can be evaluated on the basis of many interrelated criteria, including intelligibility, perceived quality, and environmental robustness. Several of these measures are clearly more significant than others, but the fact remains that speech coding algorithms do not exist in isolation from the entire voice communications system. It is therefore important that flexibility of application and robustness to severe data transmission environments be considered when evaluating speech coding methods.

## **Speech Data Compression with Sinusoidal Representations**

A variety of voice communications systems require data rate compression of the speech signal. Approaches to this problem can be separated into two categories, algorithms which make few assumptions about the nature of the input data, such as adaptive Huffman or Lempel-Ziv-Welsh (LZW) coding, and those which are optimized for the expected data, in this case speech, such as ADPCM and LPC-based coders. The first type of compression technique reduces the memory storage requirements with no information loss, while the second type involves some reduction in information as compared to the original data. Both types of coders reduce the memory usage by a fixed percentage. The method presented here allows for memory storage to be balanced against speech quality.

The basic coding technology used is the Sinusoidal Transform Coding (STC) speech compression method [6, 7, 8, 9, 10]. The STC model has been shown to be effective at various discrete rates from 8 kbps to 2.4 kbps [9]. Further, this coding technology is extremely robust to manipulation in the parameter space, as shown by the previous work in STC parameter space transformations [1]. The parameter space transformation method allows the most optimum coding for this coder type to be used for a given compression ratio. The technique presented here allows compression in multiple stages of any selected size, with no interaction with the source.

## **Speech Packet Data Loss**

Telecommunications networks are becoming increasingly diverse with the introduction of new services, such as digital mobile radio, and architectures, such as ATM-based packet networks, to serve a variety of needs as shown in Figure 1. The heterogeneous nature of such networks implies that the representation which will be used for speech signals must be robust in the presence of all degradations that may be encountered in the path from source to destination. This work indicates that parametric coders using sinusoidal representations can be used in extremely harsh packet loss environments, even beyond

Heterogeneous Network Environment

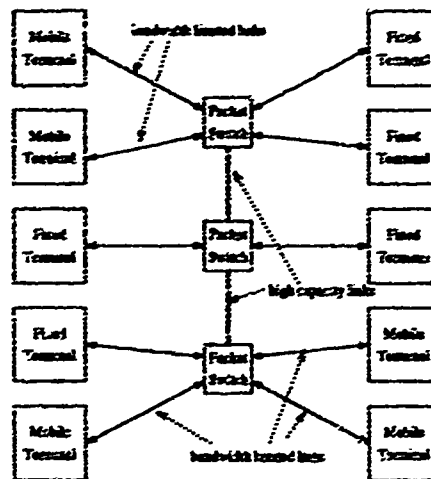


Figure 1: Heterogeneous Packet and Mobile Environment

those which degrade PCM speech to unacceptable levels. Although this work is primarily directed at data loss recovery in packet networks, the techniques described here can be used for signal reconstruction when fades or bit errors which are not corrected by conventional error control coding occur in a digital mobile radio environment.

The reconstruction of lost packets containing 64 kbps PCM and 32 kbps ADPCM speech data has been studied in relation to anticipated congestion problems in the backbone packet-based network, where excessive delay or buffer overflow can result in packet discard [2, 3, 4, 5, 11, 12, 13]. This prior work has produced methods for reconstruction of PCM packets, with acceptable uniform loss rates up to approximately 6% with 16 ms packets. Reconstruction of ADPCM using similar strategies has proved more difficult, due to the adaptive nature of the coder.

The basic coding technology used here is the Sinusoidal Transform Coding (STC) method [6, 8, 9, 10]. The STC model has been shown to be effective at various discrete rates from 8 kbps to 2.4 kbps [9]. Testing of low rate coders has indicated that long frame lengths of 30-40 ms are highly usable, which suggests that the STC parameters exhibit good time stability. Further, this coding technology is extremely robust to manipulation in

the parameter space, as shown by work in STC parameter space transformations [1]. The work presented here is primarily focused on the 4.8 kbps coder which might be applicable to digital mobile environments.

The proposed Broadband ISDN will be based on the Asynchronous Transfer Mode (ATM) standard, which defines the packet data field size to be 48 bytes. Due to delay considerations, however, it is anticipated that only one frame (20 ms for our 4.8 kbps STC) will likely be included in a single ATM packet. The results of this research are based on these assumptions.

## Results

### **Efficient, Storage Adaptive Speech Compression**

Because of the robust nature of the representation used by the STC algorithm, not only can the coder operate at various rates [9], but transformations in the parameter space can be used to change the bit rate of the coded data.

The parameter transformation technique allows the coding for all spectrum, excitation, and frame length parameters to be selected at each rate to give the optimum coding for that rate. Among the parameters which may be varied at each rate are the number of cepstral coefficients, the frequency range for coding, the method of representation (frame-fill or the actual value) of the mid-frame excitation parameters, and the number of bits used for each parameter. The method for changing the spectral representation (that is, the number of cepstral coefficients) is shown in Figure 2. The cepstral coefficients derived from the coded parameters are used to derive a spectral envelope, which is then used to generate a new set of cepstral coefficients, which can then be coded. The parameter temporal interpolation technique used for frame length changes at each stage is depicted in Figure 3. The various parameters at each new frame boundary (that is, at  $T'_{k-1}$ ,  $T'_k$ ,  $T'_{k+1}$ ), are linearly interpolated according to the distance of the new frame boundaries from the closest

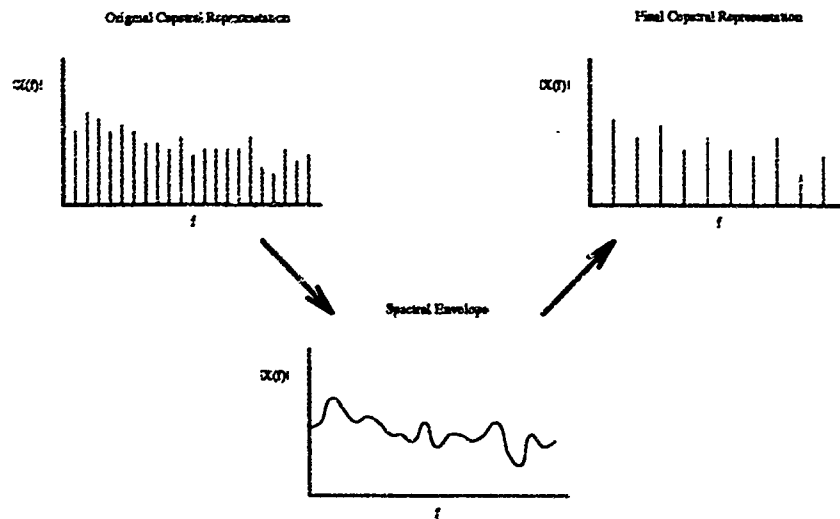


Figure 2: Technique for Changing the Spectral Representation

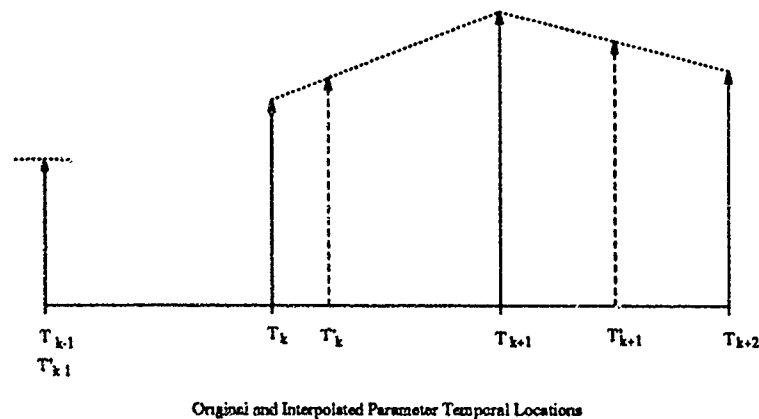


Figure 3: Interpolation of Parameters for New Frame Lengths

old edges ( $T_{k-1}, T_k, T_{k+1}, T_{k+2}$ ). Provision is also made for interpolating mid-frame and frame-fill parameters in a similar manner.

The method presented above is ideal for many applications. A typical application for this multi-level compression technique, a voice mail system, was created to test the overall concept. The structure of the prototype voice mail system is illustrated in Figures 4 and 5. The mail is generated by a sender and transmitted to the recipient via a computer network, in this case via Ethernet using the Sun Microsystems Remote Procedure Call (RPC) protocol, which in turn uses the standard TCP/IP protocol stack. The storage system processes the

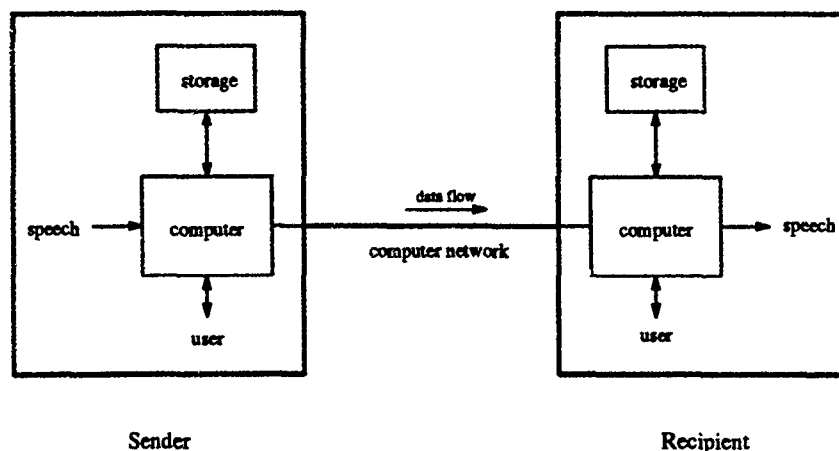


Figure 4: Voice Mail Application

arriving packets to insure that message consistency is maintained. The speech data is then passed to the STC analyzer to code the data in the STC format, if it is not already using that representation (this initial encoding provides a 4-to-1 compression over the standard 64 kbps  $\mu$ -law representation). After the initial compression, the data is stored on disk. The storage management logic then checks the storage space remaining available, and performs transformations as necessary. The memory management control logic is shown in Figure 6. The simple scheme implemented in the prototype system reduces the effective bit rate of all of the stored messages when a specified threshold is reached. In this manner, the total memory usage is reduced, with corresponding quality reductions in all of the messages. More advanced control mechanisms, such as transformation of previously checked or old messages before new messages, can be applied in a similar manner.

In order to verify that the transformation was producing high quality speech, intelligibility tests were performed. The 2.4 kbps test data was generated by performing the initial STC analysis at a 16 kbps rate with a 10 ms frame size. The rate was then reduced via the transformation to 8 kbps with a 20 ms frame length, then to 4.8 kbps with a 20 ms frame, and finally to 2.4 kbps with a 30 ms frame length. The results are summarized in Table 1. Most listeners find the the perceived speech quality to be good, only slightly less pleasing

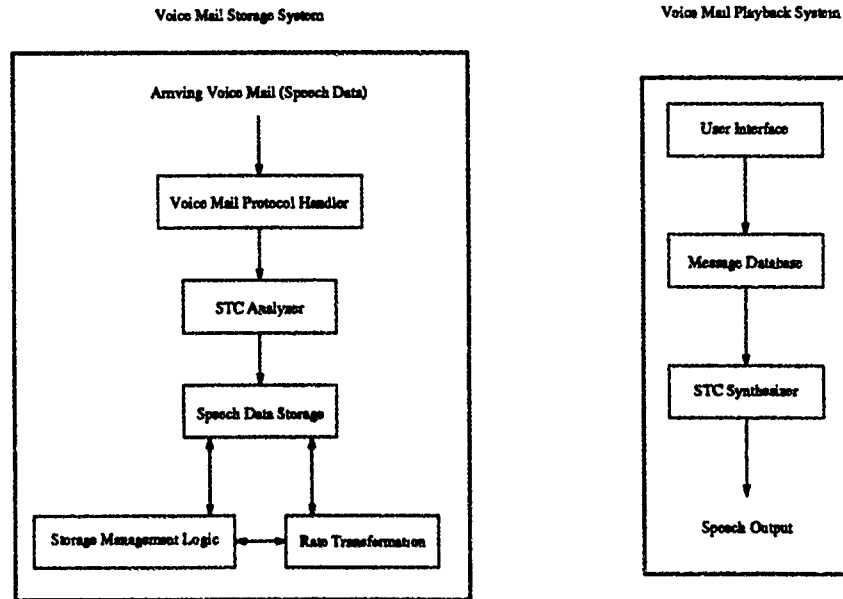


Figure 5: Voice Mail Processing System

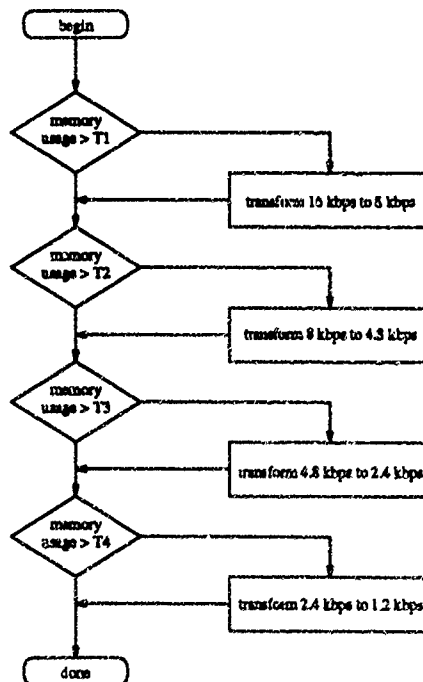


Figure 6: Message Compression Logic



Table 1: DRT Test Results

Speaker(s)	Results
JE	91.93
CH	95.44
RH	91.67
Total for 3 Males	93.01

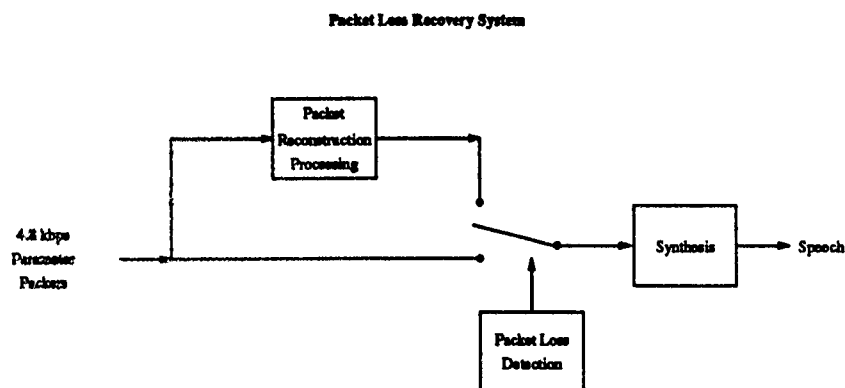


Figure 7: Packet Loss Recovery/Reconstruction System

than the untransformed 2.4 kbps STC.

### STC Packet Loss Reconstruction

Missing STC parameter packets can be reconstructed using the system shown in Figure 7. The most straight-forward method is to use the parameters received in the previous packet (parameter freezing), as depicted in Figure 8. Note that this method entails no additional delay beyond the usual analysis/synthesis time. Better performance can be attained by using the packet received after the missing packet and parameter interpolation procedures. This method is illustrated in Figures 9 and 10. Linear interpolation is used for the spectrum and excitation parameters when the adjacent frames are available. If more than one frame in a row is missing, the closest received frame is used.

In order to test the effectiveness of these methods, the STC has been informally tested at a variety loss rates, where the missing packets are uniformly distributed. The performance

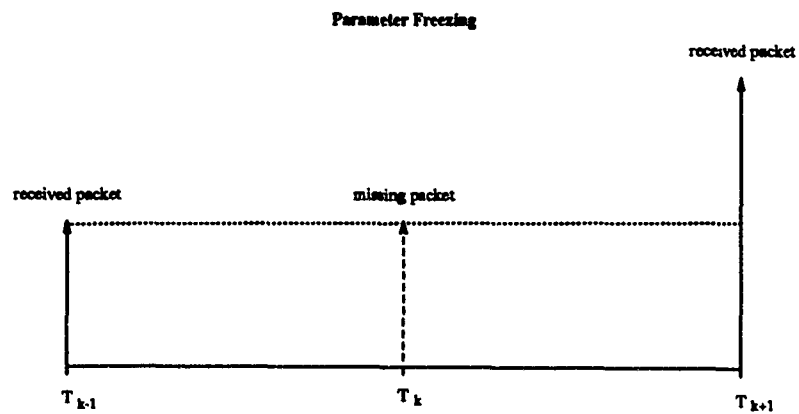


Figure 8: Parameter Freezing for STC Packet Loss Reconstruction

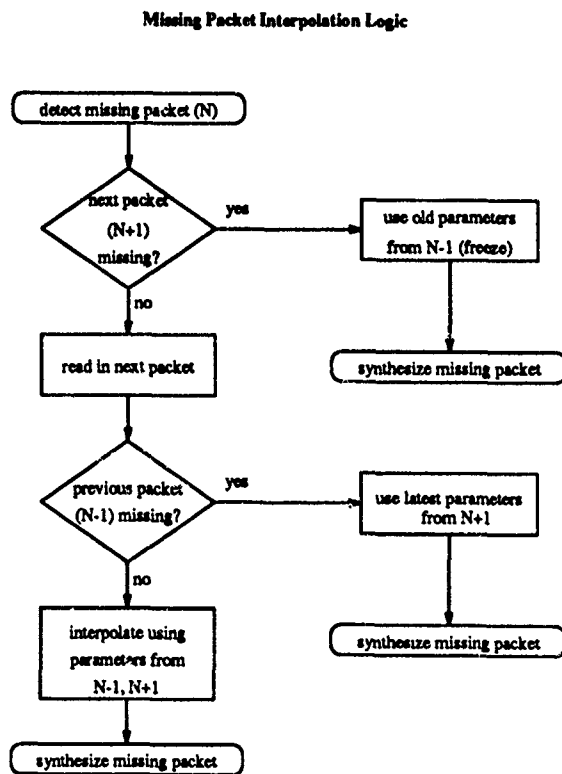


Figure 9: STC Parameter Interpolation Logic

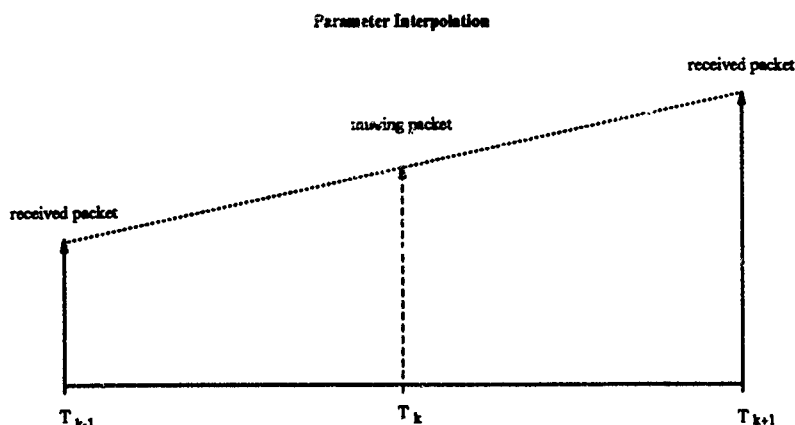


Figure 10: Parameter Interpolation for STC Packet Loss Reconstruction

of the STC in all packet loss environments was outstanding. At 12% loss rates, which were unacceptable for even the best 64 kbps PCM reconstruction methods [13], the 4.8 kbps STC was only slightly degraded when parameter freezing reconstruction was used, and the loss was almost unnoticeable when interpolation was applied. In order to verify and quantify these performance results, DAM tests are currently being performed on both the freeze and interpolation methods with 12% loss rates. The best PCM reconstruction method at this loss rate [12] is also being tested for comparison.

## Summary of Results

This research has produced several new algorithms which provide significant increases in speech coder utility. These results have been described in the following papers:

J. B. Evans and T. G. Champion, "Robust Speech Coding and Reconstruction Techniques for Heterogeneous Mobile/Packet Networks", submitted to *1992 IEEE Int. Conf. Acoust., Speech, Signal Processing*.

T. G. Champion and J. B. Evans, "A Multi-Rate STC Speech Compression Technique with Applications in Voice Mail", submitted to *1992 IEEE Int. Conf. Acoust., Speech, Signal Processing*.

Further, patent disclosures on this work have been filed with the U.S. Government:

**“A Multi-Rate STC Speech Compression Technique for Speech Data Storage”,**

**Joseph B. Evans and Terrence G. Champion.**

**“STC Speech Packet Reconstruction Techniques for Heterogeneous Mobile/Packet**

**Network Environments”, Joseph B. Evans and Terrence G. Champion.**

### **Conclusion**

The work performed under this program has resulted in new algorithms for improved capability for speech data storage and for high quality, low rate speech coding in harsh communications network environments.

A new multi-stage speech compression algorithm using parameter space transformations was presented. This new method allows speech quality to be balanced against storage requirements, while maintaining high quality for each particular rate. This algorithm allows more efficient use of resources. Specifically, in a military environment, the need for voice data storage resources (more message) may be considerably greater during times of crisis than at normal times. Some quality can now be sacrificed during critical periods, as conditions demand, while maintaining high quality when resource usage is light.

New methods of packet loss recovery for low bit rate speech systems were developed. These algorithms allow extremely high packet (frame) loss rates to be endured, albeit at some slight reduction in quality as compared to the equivalent unimpaired speech. These techniques can be applied to both packet and land mobile radio environments. When used in conjunction with traditional error control coding methods, the new algorithms can provide an exceptionally high degree of robustness in the presence of network and channel errors. The new methods provide important new capabilities for the military environment by vastly increasing the usability of STC-based voice communications in harsh jamming or transmission channel noise environments.

Both of these developments exploit the parameter space robustness of the original STC

representation. Future research will undoubtedly uncover additional applications arising from this characteristic of the STC.

## References

- [1] T. G. Champion. Theory of parameter space transformation techniques. Tech. rep., Rome Laboratories, to be published.
- [2] D. J. Goodman, G. B. Lockhart, O. J. Wasem, and W. C. Wong. Waveform substitution techniques for recovering missing speech segments in packet voice communications. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(6):1440–1448, Dec 1986.
- [3] J. G. Gruber and L. Strawczynski. Subjective effects of variable delay and speech clipping in dynamically managed voice systems. *IEEE Trans. Comm.*, COM-33(8):801–808, Aug 1985.
- [4] O. G. Jaffe. Reconstruction of missing packets of PCM and ADPCM encoded speech. Master's thesis, M.I.T., June 1986.
- [5] N. S. Jayant and S. W. Christensen. Effects of packet losses in waveform coded speech and improvements due to an odd-even sample-interpolation procedure. *IEEE Trans. Comm.*, COM-29(2):101–109, Feb 1981.
- [6] R. J. McAulay and T. G. Champion. Improved interoperable 2.4 kb/s lpc using sinusoidal transform coder techniques. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 641–643, 1990.
- [7] R. J. McAulay and T. F. Quatieri. Magnitude-only reconstruction using a sinusoidal speech model. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 27.6.1–27.6.4, 1984.

- [8] R. J. McAulay and T. F. Quatieri. Mid-rate speech coding based on a sinusoidal representation of speech. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 945–948, 1985.
- [9] R. J. McAulay and T. F. Quatieri. Multirate sinusoidal transform coding at rates from 2.4 kb/s to 8 kb/s. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1645–1648, 1986.
- [10] R. J. McAulay and T. F. Quatieri. Speech analysis-synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754, Aug 1986.
- [11] D. Petr, L. DaSilva, and V. Frost. Priority discarding of speech in integrated packet networks. *IEEE Journ. Select. Areas Commun.*, SAC-7(5):644–656, June 1989.
- [12] R. A. Valenzuela and C. N. Animalu. A new voice-packet reconstruction technique. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1334–1336, 1989.
- [13] O. J. Wasem, D. J. Goodman, C. A. Dvorak, and H. G. Page. The effect of waveform substitution on the quality of PCM packet communications. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-36(3):342–348, March 1988.

# A GENERAL ONE-DIMENSIONAL III-V HETEROJUNCTION DEVICE SIMULATOR

Dr. Ronnie E. Owens, Assistant Professor

## 1 Introduction

This report describes a general one-dimensional III-V heterojunction device simulation program which was developed by the author during the 1991 ten week summer appointment with Rome Laboratories at Hanscom Air Force Base. Research on heterojunction devices, as well as the related materials research, are important ongoing efforts which will lead to high-speed digital and high-frequency analog circuitry far outstripping the performance of silicon. In addition, almost all optoelectronic devices in use today are based on the III-V heterojunction technologies. Modeling is an especially important part of this research and can be crucial in device design and analysis. The program described herein represents an effort to provide a tool which will be useful to the device designer and analyst. The program is capable of simulating heterojunction devices composed of III-V ternary materials with compositional grading, impurity grading, optical excitation, and a general trap model intended for use in modeling the InP/oxide interface. Significant effort was expended in providing a convenient user interface to the program. In the succeeding sections the following topics are discussed: the physics background of the simulation program, a description of the use and internal structure of the program, some examples including simulations

$x$	position in the device (perpendicular to layers)
$N_c$	conduction band density of states
$N_v$	valence band density of states
$F_{\frac{1}{2}}$	Fermi-Dirac integral of order one-half
$V$	electrostatic potential
$\chi$	electron affinity
$\phi_n(x)$	quasi-Fermi level for electrons
$\phi_p(x)$	quasi-Fermi level for holes
$n$	electron density
$p$	hole density
$g$	degeneracy factor
$N_D$	donor density
$N_A$	acceptor density
$E_{DB}$	donor binding energy (i.e. relative to conduction band edge)
$E_{AB}$	acceptor binding energy (i.e. relative to valence band edge)
$\epsilon$	permittivity
$\mu_n, \mu_p$	electron and hole mobility
$\tau_n, \tau_p$	electron and hole recombination lifetimes
$J_n, J_p$	electron and hole current density
$U_n, U_p$	electron and hole recombination currents
$G$	optical generation rate
$h\nu$	incident radiation energy

Table 1: List of Symbols.

of an InP/InAlAs power MISFET, and some suggestions for future improvements.

## 2 Background

This section details the physics background for the simulator. Table 1 provides a list of symbols for this section. The basic theoretical framework for the equations is based on position-dependent band structure transport model presented by Marshak [1]. This approach references the electrostatic potential to the vacuum level. The vacuum level merely provides a convenient continuous reference level. The electron affinity is assumed to be a phenomenological parameter and is chosen so as to yield the correct band offsets between different materials. This model is essentially the Anderson model which, although suffering from criticism in recent years, still serves



as a convenient phenomenological starting point. The substrate (*i.e.* the last layer) is chosen as the potential reference. With these choices, the electron and hole densities in non-equilibrium can be written as

$$n = N_c F_{\frac{1}{2}} \left( \frac{V(x) + \chi(x) - \phi_n(x)}{kT} \right), \quad (1)$$

and

$$p = N_v F_{\frac{1}{2}} \left( -\frac{V(x) + \chi(x) + E_g(x) - \phi_p(x)}{kT} \right). \quad (2)$$

respectively. Given impurity densities as a function of position, ionized impurity distributions in non-equilibrium are given by

$$N_D^+ = N_D(x) \left[ 1 - \frac{1}{1 + \frac{1}{g} \exp \left( \frac{V(x) + \chi(x) + E_{DB}(x) - \phi_n(x)}{kT} \right)} \right], \quad (3)$$

and

$$N_A^- = \frac{N_A(x)}{1 + g \exp \left( -\frac{V(x) + \chi(x) + E_g(x) - E_{AB}(x) - \phi_p(x)}{kT} \right)} \quad (4)$$

for donors and acceptors respectively. Similar expressions yield ionized trap densities for donor-like traps,  $N_{DT}^+$ , and acceptor-like traps,  $N_{AT}^-$ , with quasi-Fermi levels for electrons and holes replaced with quasi-Fermi levels for traps,  $\phi_{DT}$  and  $\phi_{AT}$  [2]. The emission and capture rates of the traps are calculated using the model from [2] and the trap quasi-Fermi level is updated at each timestep. These terms are the source terms for the Poisson equation which, in one dimension, may be written

$$\frac{d}{dx} \left( \epsilon(x) \frac{dV(x)}{dx} \right) = n - p + N_A^- - N_D^+ + N_{AT}^- - N_{DT}^+. \quad (5)$$

Note that writing the equation in this form automatically handles boundary conditions at interfaces between materials with differing permittivities.

The electron and hole currents may be written, in general, as

$$J_n = q\mu_n(x)n(x)\frac{d\phi_n}{dx}, \quad (6)$$

and

$$J_p = q\mu_p(x)p(x)\frac{d\phi_p}{dx}. \quad (7)$$

In this form, heterojunctions are handled naturally in the drift-diffusion approximation, with additional currents resulting from position dependent density of states and electron affinity. This is thus a more convenient form for the current density than summing a drift term and a diffusion term. The mobility is assumed to be given by a low-field, low-doping limit multiplied by terms which account for impurity scattering, field dependence and temperature dependence. The parameters which describe these effects are contained in a database of materials constants and are accessed when the program initializes. Values of mobility are also interpolated for ternary materials using the empirical expressions given by Marandet et al. [3].

Additional currents also occur locally through recombination and generation. Carrier recombination is assumed to be given by the Shockley-Read-Hall statistics, viz.,

$$U = \frac{1}{\tau_n} \frac{pn - n_i^2}{n + p + 2n_i \cosh\left(\frac{E_t - E_i}{kT}\right)}, \quad (8)$$

with hole recombination assumed to be given by the same expression. The recombination lifetimes  $\tau_n$  and  $\tau_p$  are given by empirical expressions which account for variations due to doping [4]. Carrier generation may occur through optical excitation and is given at each point by [5, 6]

$$G = \Phi_{op}\alpha(x, h\nu)e^{-\alpha(x, h\nu)x}. \quad (9)$$

in which it is assumed that light of a single input frequency is incident. The absorption coefficient as a function of position and energy is given by [7]

$$\alpha(x, h\nu) = \frac{C(x)}{\omega} \sqrt{\omega - \omega_g}. \quad (10)$$

for frequencies of excitation above the direct bandgap frequency, below which the absorption coefficient is assumed to be zero. In [7] Bennett and Soref report values of  $C$  for various binary III-V semiconductors. Values for the ternary semiconductors used in this work were interpolated from these.

The electron and hole densities develop with time according to the current continuity equations given, in one dimension, by

$$\frac{\partial n}{\partial t} = \frac{1}{q} \frac{dJ_n}{dx} + G_n - U_n \quad (11)$$

for electrons and

$$\frac{\partial p}{\partial t} = -\frac{1}{q} \frac{dJ_p}{dx} + G_p - U_p \quad (12)$$

for holes.

The numerical solution of this system of equations yields the potential and quasi-Fermi levels for electrons and holes at all points in the device. With these many parameters of interest may be calculated. Terminal currents may be derived, two-dimensional charge densities at interfaces may be calculated, and the differential capacitance can be deduced by calculating the difference in the net charge between two runs.

### 3 Program Description and Use

The program, which is called `hj1d`, has more or less been implemented in standard FORTRAN 77. Some aspects of the program are Unix-specific and the best environment is a Sun/Sparc Unix workstation. Minor modifications may be made to the code that will enable it to run on other platforms. The program executable file should be saved in some location that is accessible by all those who wish to execute it and the full directory name of this location should be added to the shell variable `path`.

The equations of the previous section are all solved using a finite difference method with a non-uniform grid. The grid is automatically generated and care is taken to assure that the mesh spacing changes no more than a factor of 2 between successive points to preserve the numerical accuracy of the approximation. A maximum of three hundred mesh points are allowed in the simulation. An error message is generated if this limit is exceeded. The Poisson equation is solved using the simple algorithm presented by Mayergoyz [8]. This method basically discretizes Equation 5 according to the finite differences prescription and solves the resulting system of non-linear equations using a relaxation technique. In [8] it was demonstrated that this method is unconditionally convergent. Convergence is generally indicated if the maximum change in voltage is less than  $10^{-6}$  kT in a given iteration.

The solution of the current continuity equation is somewhat more troublesome. The discrete form of the divergence results in basically a sum of currents into and out of meshpoints. This essentially means evaluating the currents between meshpoints and this requires an estimate of the carrier concentrations between meshpoints. Us-

ing a simple linear average of carrier concentrations yields an unstable system of equations. Scharfetter and Gummel [9] solved this problem by expanding the carrier concentration between meshpoints so as to arrive at a constant current. This method provides a convergent algorithm in most cases. Again convergence is indicated by a maximum change in quasi-Fermi level for electrons and holes of less than  $10^{-6}$  kT.

The overall solution for steady state is arrived at by successively solving the Poisson equation and current continuity equation in sequence. During the Poisson solution, the quasi-Fermi levels are assumed to be held fixed, while during the current continuity solution the potential is assumed to be held fixed. This proceeds back and forth until self-consistency is achieved. For a transient solution, data from a previous run is used with the time derivatives in Equations (11) and (12) approximated by a backward difference approach. The simulation solves the Poisson and current continuity equations in sequence until the discrete form of (11) and (12) is satisfied. The solution of the current continuity equation is guaranteed to converge with this approach. It has not yet been proven whether the Poisson-current continuity system is convergent using this approach. The program will exit and print a summary of results if both the Poisson and current-continuity solvers have converged.

This program incorporates the ability to model a single trap level within the band gap. The capture cross section and areal density of interface traps may be entered at any point in the device. The two-dimensional density of traps is assumed to be spread out over a user-specified layer thickness which may be as thin as 1 Å. Thus these traps are ostensibly treated as a three-dimensional density internal to the program. The emission and capture rates of electrons and holes to both bands are calculated and

```

Ohmic Contact Device Structure
Temperature
300.0
Device Area (m^2)
100e-12
Nsrf Esrfb Phib
cont1 type=ohmic
cont2 type=ohmic
Layer Descriptions
Mat.  X      Grade  Th(A) Nd      Grade  Edb  Na  Grade  Eab  Max Mesh
algaas x=0.30 th=360 nd=1.0e17 edb=0.006 maxmsh=40
algaas x=0.30 th=60  nd=1.0e17 edb=0.006 maxmsh=10
algaas x=0.0 th=5000 nd=1.0e17 edb=0.006 maxmsh=1000

```

Figure 1: Device file for an AlGaAs/GaAs Heterojunction

the quasi-Fermi levels for the traps are updated at each timestep. Thus the transient occupation of trap states at an interface may be accounted for and the concomitant effects on the characteristics of an InP MIS capacitor will be presented in the next section.

The user input to the program is an ASCII text device file which provides a description of the device to be simulated. The general format of the file is shown in Figure 1. The first two lines are comments, followed by a line with the device simulation temperature. Next is another comment line and the device area. Another comment line follows and the next two lines are the top and bottom contact descriptions. If `type=ohmic` appears on the line, the program assumes no built-in barrier exists and that equilibrium exists at the contact meshpoint. Voltages are applied by changing the quasi-Fermi level and conduction band edge and leaving these as

<b>x</b>	Mole fraction (ternaries only)	0
<b>xgrda</b>	Parabolic material grade factor ( $\frac{1}{A^2}$ )	0
<b>xgrdb</b>	Linear material grade factor ( $\frac{1}{A}$ )	0
<b>th</b>	Layer Thickness (Å)	1000
<b>nd</b>	Donor density (cm <sup>-3</sup> )	0
<b>ndgrda</b>	Parabolic donor grade factor ( $\frac{1}{A^2}$ )	0
<b>ndgrdb</b>	Linear donor grade factor ( $\frac{1}{A}$ )	0
<b>edb</b>	Donor binding energy (eV)	0.006
<b>na</b>	Acceptor density (cm <sup>-3</sup> )	0
<b>nagrda</b>	Parabolic acceptor grade factor ( $\frac{1}{A^2}$ )	0
<b>nagrdb</b>	Linear acceptor grade factor ( $\frac{1}{A}$ )	0
<b>eab</b>	Acceptor binding energy (eV)	0.028
<b>maxmsh</b>	Maximum mesh spacing for layer (Å)	1000
<b>ndt</b>	Donor-like trap density (cm <sup>-3</sup> )	0
<b>edt</b>	Donor-like trap binding energy (eV)	0
<b>sign</b>	Donor-like trap capture cross section (cm <sup>-2</sup> )	0
<b>nat</b>	Acceptor-like trap density (cm <sup>-3</sup> )	0
<b>eat</b>	Acceptor-like trap binding energy (eV)	0
<b>sigp</b>	Acceptor-like trap capture cross section (cm <sup>-2</sup> )	0
<b>munmod</b>	Electron mobility model (see text)	0
<b>mupmod</b>	Hole mobility model (see text)	0

Table 2: Table of layer description paramters and default values.

boundary conditions for the current continuity and Poisson solvers respectively. If **type=schottky** appears, then the **phib=xx** is scanned and the quasi-Fermi level is assumed to be pinned at the value given (eV) below the conduction band edge. Next follows two comment lines and the layer descriptions. Each layer description line begins with the name of the material and is followed by a list of keywords with an equal sign and a value (no spaces are allowed except between parameters). The material is the only position dependent parameter in the layer description. Table 2 describes the parameters which appear on the line of the layer description and the default value of each parameter. The materials which are presently included in the database are SiO<sub>2</sub>, InP, Al<sub>x</sub>Ga<sub>1-x</sub>As, In<sub>x</sub>Ga<sub>1-x</sub>As, and In<sub>x</sub>Al<sub>1-x</sub>As. The device file should be in the current directory and it is generally recommended that a separate sub-directory

be created for each unique device before running the program. If any changes are made to the device file which result in a different mesh, (*e.g.* additional layers or thickness changes, maximum mesh spacing changes, etc.) then a new run should be made. This is defined here as a "cold" start.

All the materials files and files which provide tabulated values of the Fermi-Dirac integrals which appear in the equations should be saved in a directory which is pointed at by the environment variable `DATAPATH`. This environment variable should be set (using the Unix command `setenv`) in the user's `.cshrc` file.

Assuming at this point that the user has generated a device file in a sub-directory, the program may be invoked by "`hjid device-file args`". The device file name will be assumed to end in `.dev` if an extension is left off. Table 3 presents a list of command line arguments that may be used with the program. Note that the equal sign and a number are required for all command-line arguments except the device file name and `cold`.

A typical run will now be described. Assume that the device file of Figure 1 has been entered in a file which has been named `contact.dev`. The procedure for generating the steady-state I-V characteristics from 0 to 1.0 volts of applied bias will now be described. First, a cold run must be made to arrive at the equilibrium solution of the Poisson equation. This is accomplished by issuing the command "`hjid contact cold`". This generates a number of raw and processed data files. The raw files may essentially be ignored, while the processed files may be used to generate graphical output of the simulation results. The processed data files are `band.dat`, `pn.dat`, `recomb.dat`, `ionized.dat`, `jn.dat`, and `jp.dat`. These are respectively,



<b>cold</b>	Indication of an equilibrium (0 bias) run	no
<b>imain=</b>	Number of main iterations to perform	1000
<b>ipoi=</b>	Number of Poisson iterations per main loop	1
<b>icur=</b>	Number of Current iterations per main loop	1
<b>popt=</b>	Optical power incident (W)	0
<b>eopt=</b>	Energy of incident photons (eV)	0
<b>delt=</b>	Timestep; simulated time increment beyond saved run (s)	1e12
<b>vg=</b>	Voltage to apply to surface contact (V)	Previous Voltage
<b>gamma=</b>	Over-Relaxation parameter for iterative solution	1

Table 3: Table of *command line arguments* and default values.

band diagram, electron and hole concentration, recombination rate, ionized impurity density, electron current density, and hole current density vs. position in the device. These files contain  $x - y$  coordinate pairs describing the aforementioned data. Data will also appear on the screen describing the progress of the simulation. A summary of the device will be printed out and an iteration by iteration account of the simulation is printed. A summary of results is printed after the simulation converges. The summary lists terminal currents, capacitances, and areal electron and hole densities. This summary may be saved to a file by redirecting the output of the program to a file (e.g. `hj1d contact cold >out` will redirect the output to a file called out). Next, a "saveit" command should be issued (this is a shell script which should be installed in the same directory as `hj1d`). This command saves the most recently

generated solution for future use. Next, the command "hj1d contact vg=0.02" is issued which will simulate the device with an applied bias of .02 V. Again the output of the simulation may be saved to a file for later use.

Following the convergence of this run, a projection technique may be used to generate the initial guess for the next run. This is calculated from the present run and the saved run. Therefore the projected run will be at .04 V. The command used is `project`. Next, another `saveit` is issued to save the present run and then the command `mvit` is issued to move the projected run into place in preparation for the next run. Finally, "hj1d contact" runs the program for on the recently projected data.

The procedure from here on up to 1 V of bias is the following:

```
% project  
% saveit  
% mvit  
% hj1d contact
```

Note that the voltage step per run is given by the choice of bias for the second run (the one following the cold start). It is possible that if the voltage step is too large that this run, or perhaps a subsequent run, will diverge. This is generally indicated by currents that are nonsensically large or by current continuity which is not satisfied.

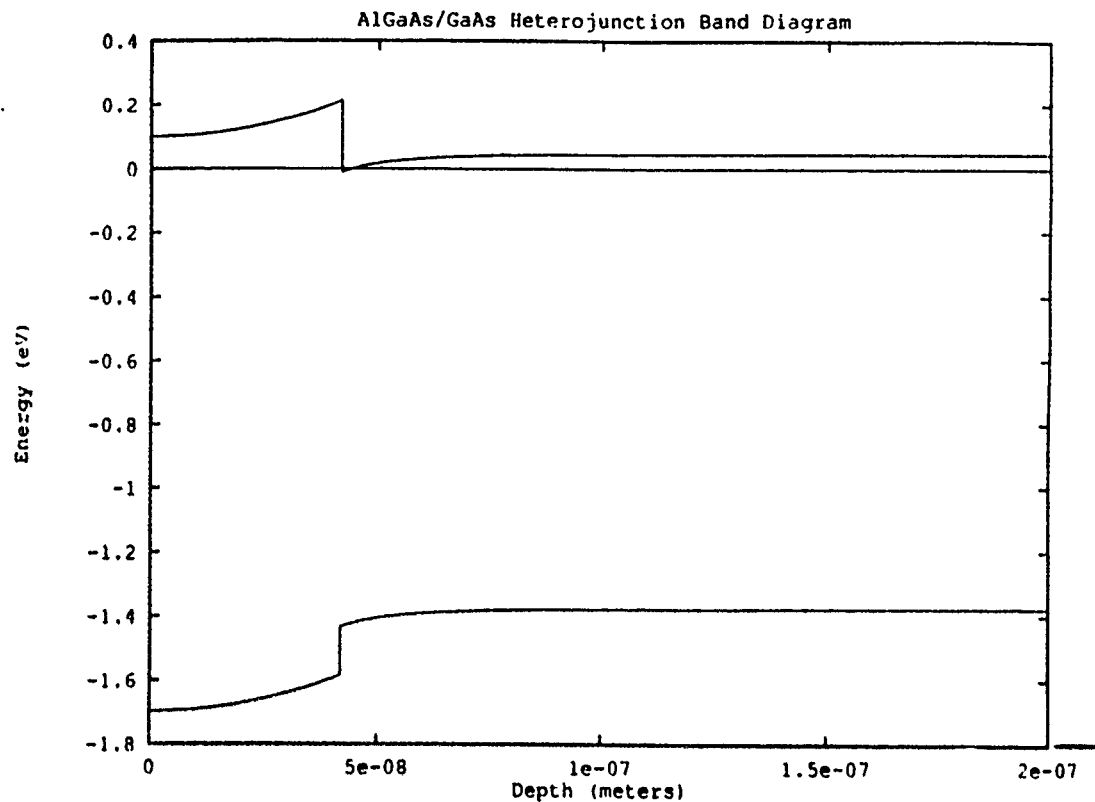


Figure 2: Band Diagram of the AlGaAs/GaAs Heterojunction.

## 4 Example Results

The results of several example devices are presented in this section. First the example run of the previous section is examined. This device is basically a heterojunction ohmic contact and would be commonly be found in a MODFET structure. Both the surface and substrate contacts are assumed to be ohmic. Since the doping level is only  $10^{17} \text{ cm}^{-3}$  The heterojunction itself dominates the I-V characteristics. Figure 2 shows the equilibrium band diagram of the structure. This result would be stored in the `band.dat` file after the cold start run. Figure 3 shows the I-V characteristics of the structure. Note the non-linearity in the characteristics and the assymetry.

To demonstrate the programs opto-electronic device capability, the simulation of an InP/InGaAs heterojunction phototransistor are presented next. Figure 4 shows

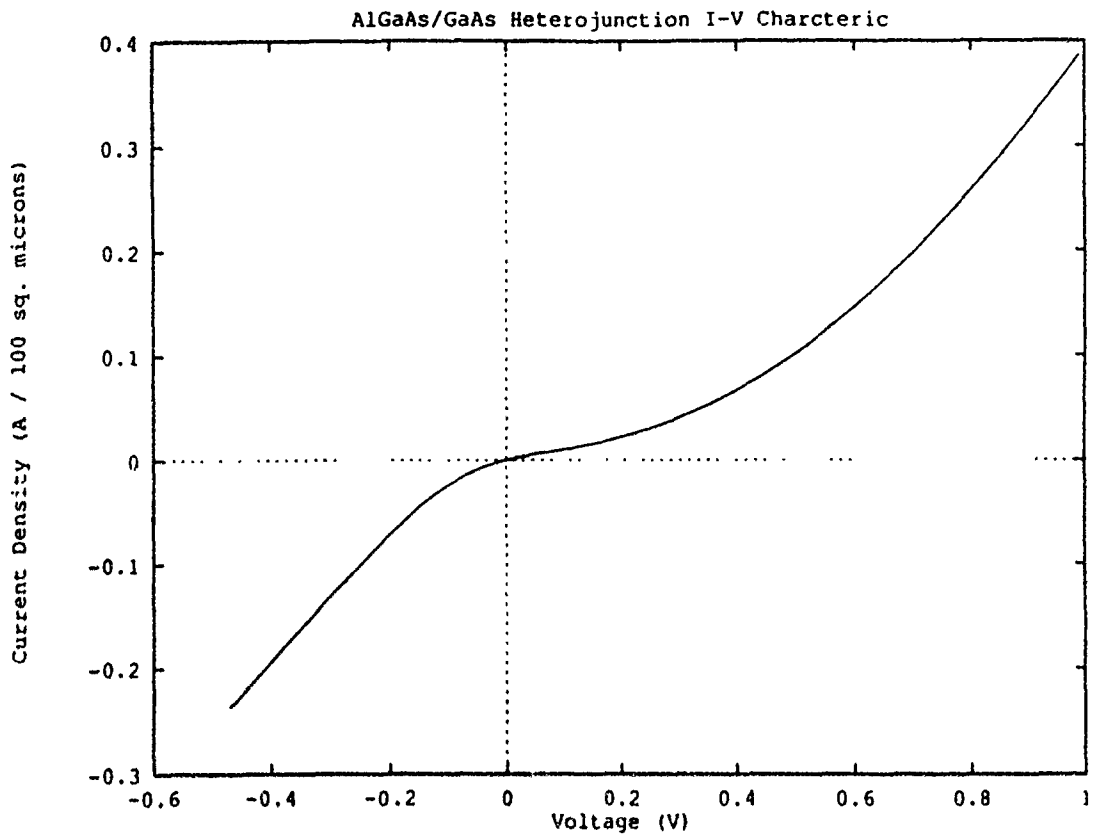


Figure 3: Current-Voltage characteristic of an AlGaAs/GaAs Heterojunction.  
InGaAs/InP Band Diagram:  $V_{ce}=1.0$ ,  $P_i=250\mu W$

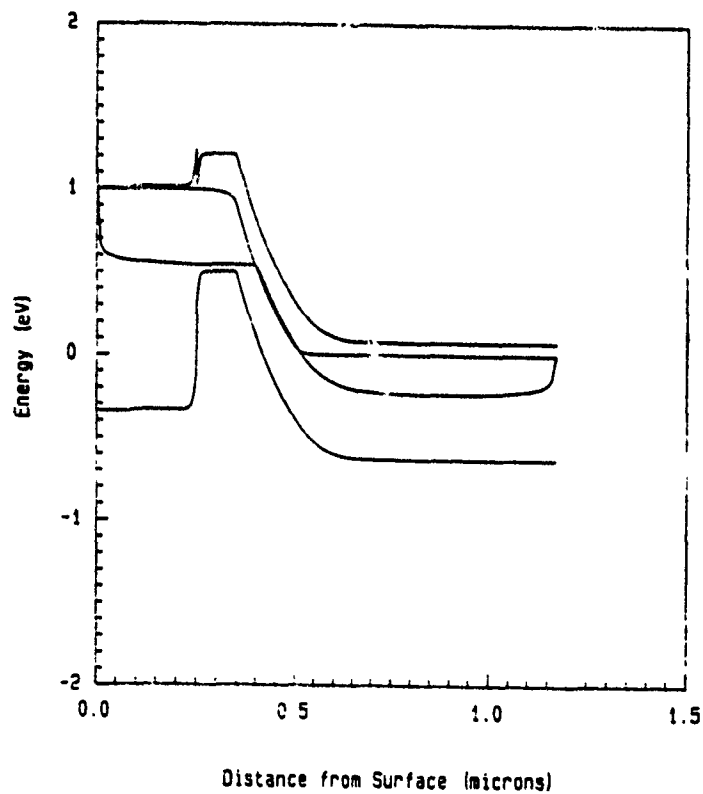


Figure 4: Band diagram of an InP/InGaAs heterojunction phototransistor.

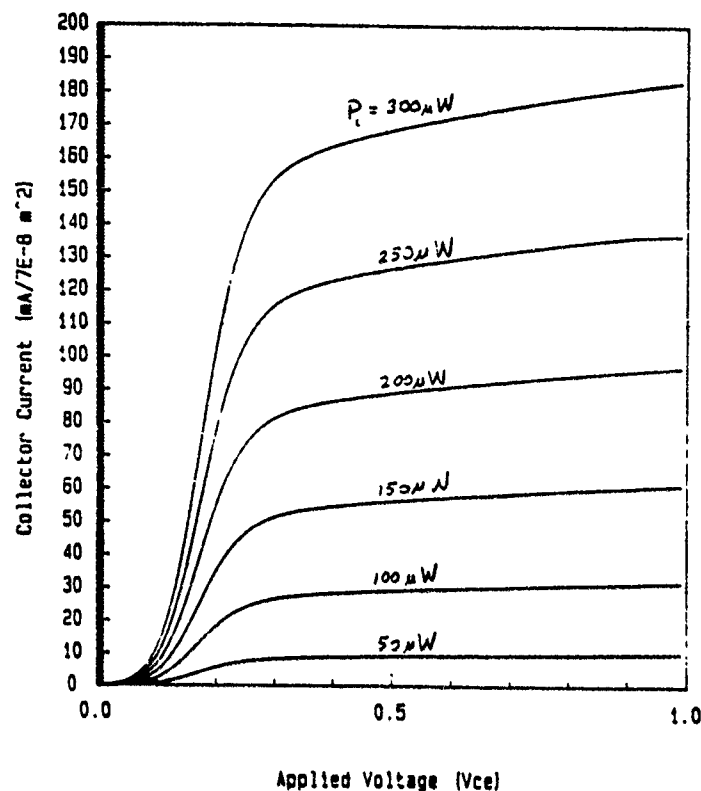


Figure 5: Common emitter I-V characteristics of InP/InGaAs heterojunction phototransistor.

the band diagram of an InP/InGaAs HPT under a bias of 1 V and an incident optical power of  $250\mu\text{W}$ . The device area is  $7 \times 10^{-8} \text{ m}^{-2}$ . The common-emitter characteristics of this device for different values of optical input power are given in Figure 5. The incident photons have an energy of 1 eV and at this energy pass through the transparent InP emitter layer and are absorbed in the base and collector layers. This device has a base width of  $0.1\mu\text{m}$ . Figure 6 demonstrates that the gain of the device is a weak function of the optical input intensity.

Finally, and the actual main thrust of the project, the results of simulations of an  $\text{SiO}_2/\text{InP}/\text{InAlAs}$  MIS capacitor are presented. FETs made from this structure show great promise in high frequency applications. This is basically a depletion mode device and the (undoped) InAlAs layer provides a barrier which must be depleted

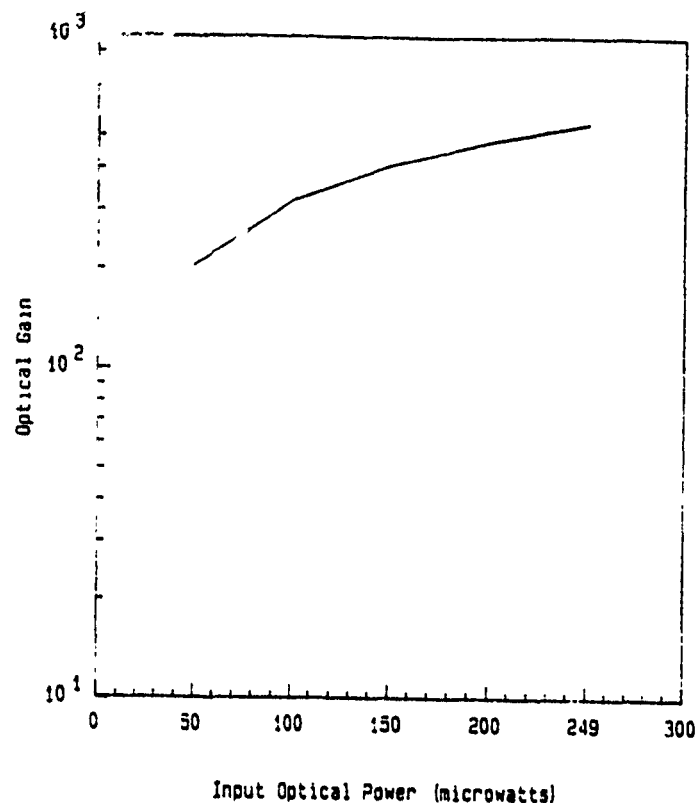


Figure 6: Gain versus input optical power of InP/InGaAs heterojunction phototransistor.

through to shut the device off. The band diagram is shown in Figure 7. The first simulations of this device were of the ideal structure, *i.e.* with no interface states. The capacitance versus voltage is shown in Figure 8 for three different cases. In the deep depletion case, a voltage ramp is applied to the gate of the device which is sufficiently fast that a hole inversion layer cannot form. The high frequency case intersperses steady state runs to adjust to a new bias with a  $1\mu\text{s}$  0.1 V ramp applied thereafter to simulate the high frequency capacitance measurement.

To demonstrate the capability of simulating interface traps in this device a series of runs was undertaken with a mid-gap donor-like trap at the  $\text{SiO}_2/\text{InP}$  interface. The density of these traps is  $10^{12} \text{ cm}^{-2}$ . This density of traps will cause a distortion of the CV curve at a gate bias at which the interface quasi-Fermi level passes through

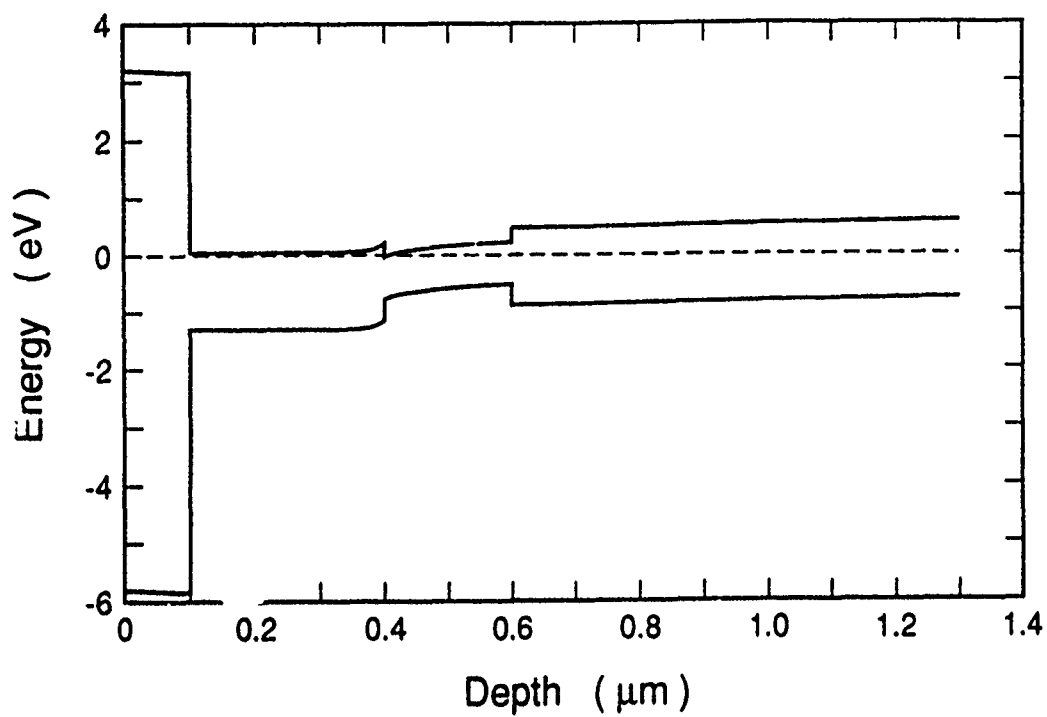


Figure 7: InP/InAlAs power MISFET band diagram.

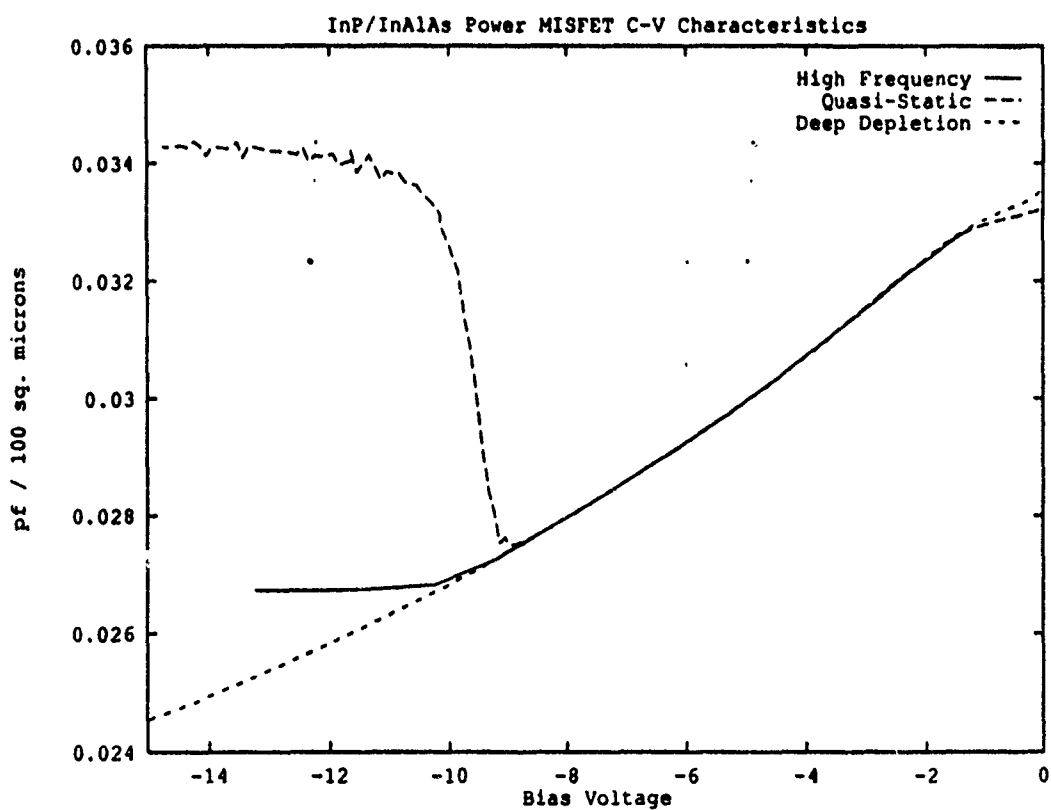


Figure 8: Ideal InP/InAlAs MIS capacitor CV curves.

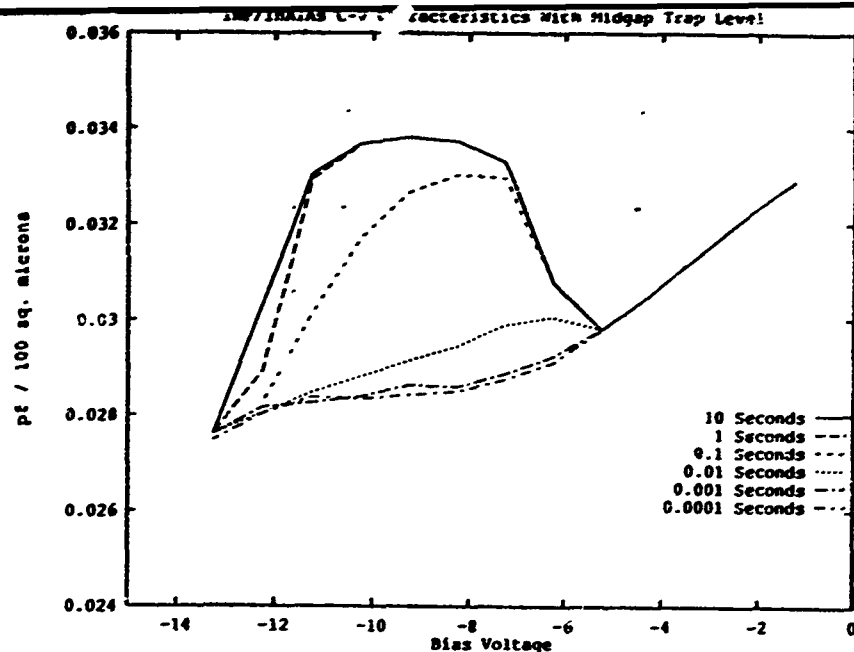


Figure 9: InP/InAlAs MIS capacitor CV curves with  $10^{12} \text{ cm}^{-2}$  donor-like traps at mid-gap. Runs with different capacitance signal ramp rates are also shown, demonstrating the transient capability.

the trap level. This distortion is shown in Figure 9 in which it may be noted that the capacitance rises to the oxide capacitance as the traps begin to empty. Different measurement times for the capacitance sensing signal are given in the figure showing the transient trap occupation for finite measurement times. This indicates that under certain measurement conditions this structure will show strong hysteresis effects as the capacitance is measured.

## 5 Future Work

The simulation program described in this report is by no means complete. The next obvious step is to implement a full blown interface trap model which will allow a distribution of traps and capture cross sections throughout the forbidden gap. This will allow modeling of the capacitor to be compared to the actual devices. This



addition to the model should be a straight-forward extension of the single trap level that is currently implemented. This model would be similar to the one presented by Hasegawa *et al.* [10, 11].

The program also contains the skeletal beginnings of a three-valley drift-diffusion model. Fleshing this out would allow the modeling of devices in which the material grade gives rise to crossover into indirect bandgap materials. Such devices include laser structures with graded index of refractions to permit confinement of radiation.

High field effects are modeled presently through the use of three mobility models. With the three valley simulation, it may be possible to model these non-stationary effects in a more physical manner. Work is underway to investigate this possibility.

In addition to this the addition of quantum effects is also under investigation. Such effects would include carrier quantization at interfaces and tunneling transport.

Finally, the model will be extended to two dimensions. This is perhaps the most important step in modeling the power MISFET. This work would provide some insight into some of the anomalous behavior the devices have exhibited.

## References

- [1] Alan H. Marshak, "Modeling Semiconductor Devices with Position-Dependent Material Parameters," *IEEE Trans. Electron Devices*, vol. 36, no. 9, 1764, September 1989.
- [2] W. Shockley and W. T. Read, "Statistics of the Recombinations of Holes and Electrons," *Physical Review*, vol. 87, no. 5, September 1, 1952.

- [3] F. Marandet, J. Bernard, P. Garcia, J. Deforges, D. Magnant, *Solid State Electron.*, vol. 32, no. 8, 1989.
- [4] D. J. Roulston, N. D. Arora and S. G. Chamberlain, *IEEE Trans. on Electron Devices*, ED-29, Feb. 1982.
- [5] Joseph E. Sutherland and John R. "A Computer Analysis of Heterojunction and Graded Composition Solar Cells," *IEEE Trans. Electron Devices*, vol. ED-24, no. 4, April 1977.
- [6] K. Yokoyama, M. Tomizawa, H. Kanbe, T. Sudo, "A Numerical Analysis of a Heterostructure InP/InGaAs Photodiode," *IEEE Trans. Electron Devices*, vol. ED-30, no. 10, October 1983.
- [7] Brian R. Bennett and Richard A. Soref, "Electrorefraction and Electroabsorption in InP, GaAs, GaSb, InAs, and InSb," *IEEE J. Quantum Electron.*, vol. 26, no. 1, January 1990.
- [8] I. D. Mayergoyz, *J. Appl. Phys.*, vol. 59, no. 1, January 1, 1986.
- [9] D. L. Scharfetter and H. K. Gummel, *IEEE Trans. Electron Devices*, vol. ED-16, no. 1, 64, January 1969.
- [10] H. Hasegawa, L. He, H. Ohno, T. Sawada, T. Haga, *J. Vac. Sci. Technol B*, vol. 5, no. 4, July/August 1987.
- [11] L. He, H. Hasegawa, T. Sawada, H. Ohno, *Jap. J. Appl. Phys.*, vol. 27, no. 4, April 1988.

# **FDTD ANALYSIS OF THE RADIATION PROPERTIES OF A PARABOLIC CYLINDER ILLUMINATED BY A VERY SHORT PULSE**

Carey M. Rappaport  
Assistant Professor

## **ABSTRACT**

This report addresses the problem of extremely short-time pulse excitation of a parabolic reflector antenna. The pulse is selected to be short compared to the time required for it to traverse the reflector, so that only a fraction of the reflector is illuminated at any given time. Because the pulse is so short, standard frequency domain methods of analysis are impractical. Instead, the Finite Difference Time Domain method is used to analyze the wave propagation and reflection. The field distributions have been simulated for a deep parabola in both transmit and receive modes. The results indicate that commonly perceived assumptions of differential transient reflection are inaccurate.

## **INTRODUCTION**

As signal generation technology improves, the need to understand the behavior of shorter and shorter pulses becomes increasingly important. When the duration of the pulse becomes comparable to—or shorter than—the typical length scale of an antenna system, new ways of studying electromagnetic field propagation become necessary.

One important problem that has recently been studied [1,2] is the excitation of a parabolic reflector antenna by a pulse that is shorter than its focal length. The

traditional analysis method of tracing rays is brought into question since the rays do not all illuminate the reflector simultaneously.

In fact, Hansen first states [1] that the leading and trailing edges of a short pulse have different antenna patterns, leading to a differential transient radiation characteristic. The leading portion of the transmitted wave is first reflected by the reflector surface in a circle nearest the vertex while the reflector rim is unilluminated, while subsequently, the trailing pulse edge has left the vertex it still illuminates an annular region near the rim. See Figure 1. The conclusion is that the initial radiation pattern is that of a smaller diameter aperture, while the final pattern is that of an annulus with constant outer diameter.

This view was later reversed [2] with an argument that rays diffracted from the edges of the reflector leave the antenna first, before reflected rays from the vertex. The transient pattern should instead be that of an annular aperture first, followed by the gradually shrinking circle.

The object of the current study is to show that neither conclusion is accurate, but instead that the parabolic reflector has no characteristic transient behavior other than a simple aperture. Two approaches will be used: first, a geometric argument based on a limiting case, and second, a numerical study using the Finite Difference Time Domain (FDTD) method. The FDTD method is not based on extending frequency domain analysis across a broad frequency spectrum, but instead formulates the wave propagation completely in the time domain. The shorter the time pulse, the easier it is to model the field interaction.

Although it must be cautioned that using numerical methods to prove theoretical assertions is dangerous, it is the transient behavior that is of concern, and only a procedure which emphasizes the different temporal aspects of the field can be effectively employed.

## GEOMETRIC ARGUMENT FOR UNIFORM TEMPORAL BEHAVIOR

It is well known that the paraboloid of rotation transforms microwave, single frequency spherical waves centered at its focus into plane waves leaving the reflector. Geometric optics is used as a high frequency approximation, treating waves as surfaces perpendicular to idealized rays. Equal path length of every ray traced from the focus to the reflector and then to an aperture plane perpendicular to the axis of rotation ensure that a planar phase front will be formed. Thus even though different parts of the reflector are illuminated at different times, all rays arrive at an aperture plane simultaneously.

The geometric optics approximation breaks down for lower frequencies—more precise analyses, based directly on the wave equation must be used. Consider for computational simplicity a two dimensional parabolic reflector illuminated by a longitudinally polarized, uniform cylindrical wave with center at the focal point. This incident wave is given by:

$$\overline{E}_I = \hat{z} H_0^{(2)}(k\rho) \quad (1)$$

where  $H_0^{(2)}$  is the second order Hankel function and  $k$  is the wave number. The desired reflected and transmitted plane wave propagates in the  $y$ -direction, and is represented by:

$$\overline{E}_R = \hat{z} f(x) e^{-jky} \quad (2)$$

with some yet to be specified amplitude dependence  $f(x)$ .

The equation for the parabola with focal length  $f$  is:

$$y = \frac{x^2}{4f} - f \quad (3)$$

This geometry is shown in Figure 2. Dotted plane wavefronts are continued past the reflector to indicate the virtual planar source corresponding to the focal point

If the parabola is a perfect conductor, the boundary condition that tangential electric field be zero requires:

$$\bar{E} = \bar{E}_I + \bar{E}_R \quad (4)$$

on the parabola. In fact, the condition is never exactly satisfied with the desired choice of transmitted wave, but the actual error is quite small for most frequencies of interest.

To quantify the error, first asymptotically approximate the Hankel function with:

$$\begin{aligned} H_0^{(2)}(k\rho) &= \frac{e^{-jk\rho - j\pi/4}}{\sqrt{k\rho\pi/2}} \\ &= \frac{e^{-jk(y+2f) - j\pi/4}}{\sqrt{k(y+2f)\pi/2}} \end{aligned} \quad (5)$$

where the second equation specifies the cylindrical field on the parabola. Since the approximate phase dependence of the incident field of Equation (5) is the same as that of the reflected field, Equation (2), the boundary condition can be satisfied, if the amplitude of the reflected wave is set to  $f(x) = e^{j(2kf - \pi/4)} / \sqrt{k\pi(x^2/4f + f)/2}$  on the parabola. It remains to measure how good the asymptotic approximation is.

Figures 3a and 3b show the difference in phase, and the ratio of the magnitude difference to magnitude between the Hankel function and its approximation as a function of electrical radius. Note that at as little as one wavelength ( $k\rho = 6.28$ ) the phase error is about .02 radians (roughly  $1.5^\circ$ ), and the magnitude difference is one part in one thousand. The asymptotic approximation is very good. For any practical reflector antennas, the focal length would be at least tens of wavelengths. The error would indeed be negligible.

For all frequencies of interest, the parabola reflects waves as expected. For

transient illumination, especially very short pulse excitation with predominantly high frequency components, each of the separate frequencies in the pulse spectrum is reflected in the same way. Each frequency component is imaged by the reflector into a plane wave, originating at the image plane indicated on Figure 2.

Each frequency has a separate propagation constant, so it arrives at an arbitrary aperture plane with different phase and amplitude, but none of the individual components is reflected differently. The transient nature of the pulse with many frequency components shows up as a possible change in pulse shape at the aperture, but in no (high frequency) case is there any transient variation across the aperture. Of course, once the wave leaves the reflector, diffraction at the reflector rim will affect each frequency component differently, but no differently than an aperture in a conducting plane affects a normally incident plane wave. Thus there is no differential transient effect on the transmitted wave *due to the parabolic reflector*.

## THE FDTD METHOD OF WAVE ANALYSIS

To confirm the geometrical argument presented above, the transient wave behavior is now analyzed directly, using a time domain numerical method: Finite Difference Time Domain (FDTD). This method was first proposed on a cubical lattice by Yee [3], and then, more recently, popularized by others [4-6]. Other Finite Difference techniques, which make use of triangular, conformal meshes in both the time and frequency domains have been presented [7,8].

The idea behind FDTD is that Maxwell's partial differential equations are replaced by multi-dimensional centered difference equations in space and time. First, space is discretized into two sets of interlocking cubical meshes, with the cube corners of one mesh coinciding with the cube centers of the other. The magnetic flux

of Faraday's Law is solved at the center of every cube face in terms of the electric field on the edges that bound that face; and the electric flux of Ampere's Law is solved at the center of every cube face of the complimentary mesh in terms of the magnetic fields on the edges that bound it. With isotropic media the electric flux and the electric field are related by a scalar permittivity constant,  $\epsilon$ , so that knowing the first from Ampere's Law allows the computation of the magnetic flux from Faraday's Law. The field values on one mesh are computed at one-half a time step ahead of those of the other, so that the difference equation in time at one centered space point can be equated to the difference equation in space at the one centered time point of the complimentary mesh.

One important detail, which must be carefully addressed for exterior, unbounded geometries, is simulating the wave's radiation to infinity. Although the computational domain terminates at the end of the array lattice, the wave must act as if it were continuing outward. In other words, the wave must not reflect off the artificial boundary. The boundary must absorb all incident waves; hence it is referred to as an Absorbing Boundary Condition (ABC). In principle, no numerical condition can absorb waves from every incident angle, but there have been several ABC's proposed which absorb those near normal incidence [10-13] which makes use of pseudo-differential annihilation operators. A new ABC method which is based on simulated anechoic chamber absorber has recently been presented [14]. This method appears to be preferable for very large scatterers and higher frequencies.

## SIMULATION RESULTS AND ANALYSIS

For the parabolic reflection problem, a two-dimensional FDTD method was implemented on a vector-processor enhanced Digital Equipment Corporation VAX



9000 computer. The ABC used is the Engquist-Majda [11] planar condition on forward, backward, right and left edges of the computational domain. The parabolic segment is chosen to be four times as wide as deep ( $F/D = 0.25$ ). This is typical of deep reflector antennas, with focus in the plane of the reflector rim.

The Courant Condition, that the time step times multiplied by phase velocity of the wave must be less than the spatial step is satisfied by selecting:  $r = c\Delta t/\Delta x = 0.5$ .

Three wave reflection cases are considered: 1) a received plane wave with Gaussian pulse shape; 2) a transmitted cylindrical wave with Gaussian amplitude shape; and 3) a received plane wave with a Gaussian pulse shape modulated with a high frequency carrier signal.

The Gaussian pulse with peak at  $x = x_0$  when  $t = 0$  and  $1/e$  level width of  $2w$  is represented in space and time as:

$$\overline{E} = \hat{z}e^{-(\frac{x-x_0-ct}{w})^2} \quad (6)$$

for electric field, and for magnetic field:

$$\overline{H} = -\hat{y}e^{-(\frac{x-x_0-ct}{w})^2}/\eta \quad (7)$$

For the first test case, the parabolic focal length is selected to be 100 mesh units,  $w = 20$  and  $x_0 = 40$  and the parabola vertex is placed at  $x = 180$ . To ensure that no left and right edge effects, it was necessary to choose the mesh width to be 2.5 times larger than the length. That is, the mesh is 200 by 500 mesh units. Figure 4a shows the initial pulse on a 50 by 50 point grid, where every fourth point in  $x$  and every tenth point in  $y$  is shown. The pulse peak first encounters the rim of the parabola at  $x = 80$  after travelling 40 spatial steps, or, since  $r = 0.5$ , in 80 time steps.

Figure 4b shows the beginning of the reflection at the reflector rim at  $t = 100$ . Note that the reflected wave is negative, as expected. For time samples  $t = 200$ , 300, 400, and 500 the wave continues to reflect and then focus at the focal point, as shown in Figures 4c-4f. The pulse peak passes the focus after 480 time steps. It is clear in Figure 4f, with greater amplitude scale than the previous time samples that the wave is focusing with greatest intensity at the parabola focus.

For the second case, a cylindrical wave is specified for a transmitted field. Using the asymptotic approximation for the Hankel function given above—along with a Gaussian amplitude envelope in radius—gives the transient pulse initial conditions:

$$\overline{E} = \hat{z} \frac{e^{-(\frac{\rho-\rho_0}{w})^2}}{\sqrt{\rho/\rho_0}} \quad (8)$$

and for magnetic field:

$$\overline{H} = \hat{\phi} \frac{e^{-(\frac{\rho-\rho_0}{w})^2}}{\eta \sqrt{\rho/\rho_0}} \quad (9)$$

Note that these initial conditions are approximations to an actual cylindrical wave at a given time sample, but that these initial conditions do not have to be exact for a cylindrical wave to propagate. Once the algorithm starts, the laws of wave propagation are strictly followed. Any imperfections in the initial conditions will lead to an inward travelling wave, or a change in the outward cylindrical wave shape.

Figure 5a shows the initial wave after 50 time steps, before it has encountered the reflector. The mesh is 500 by 500 points, with every tenth value displayed, and with  $\rho_0 = 55$  and  $w = 15$ . The wave spreads, reflects off the parabola, and forms a plane wave, as indicated in Figures 5b-5d, corresponding to time samples at  $t = 150$ , 350, and 550. The reflected wave is quite well-formed with the expected amplitude taper towards the edges, but with perfectly uniform phase. In Figure 5c, the peak of the reflected wave has just past the plane of the reflector rim. It is a relatively level planar phase front. In the next Figure, 5d, the plane wave is

beginning to spread, and its amplitude has more of a taper. Also visible in Figure 5d is the absorbing nature of the back boundary. The advancing cylindrical wave appears unaffected by the end of the mesh.

The third case studied is a high frequency modulated Gaussian pulse. The initial pulse condition is the same as in case 1, Equations (6,7) with a multiplication by  $\cos kx$ . Figures 6a-6f show the time evolution across a 500 by 500 point mesh, with every tenth value displayed. Figure 6a shows the eight high frequency wave peaks within the Gaussian pulse, just as it encounters the parabola at  $t = 200$ . With this choice of carrier wavelength, the reflector has a width of  $50\lambda$ . The reflections from the parabola are quite visible in Figure 6b at  $t = 300$ . In Figure 6d, at  $t = 480$ , the pulse peak is at the focus, with amplitude of about 5.5. The focal point is clearly visible, highly resolved with this high frequency signal. Figures 6e and 6f show the wave advancing past the focal point, decreasing in amplitude and spreading transversely while retaining its modulation in the propagation direction.

## CONCLUSIONS

An analysis has been made of the reflection characteristics of parabolic reflectors to short pulses. Contrary to published theory, the parabola does not produce any unexpected transient effects beyond the standard dispersive effects of free-space propagation of a short pulse.

The FDTD method was used to analyze the transient behavior of the two-dimensional parabolic reflector. In all three of the numerical cases considered, the wave behaved as expected, focusing to high intensity at the parabola focal point, or forming a planar phase front. The reflector seems to have no differential effects on the various time levels of the pulse—the reflection of the initial, leading edge of the

pulse joins correctly with the subsequently reflected trailing portions of the pulse.

There will be the same transient behavior of a wave pulse for parabolic reflection as for propagation in free space, but the parabola itself does not cause any transient non-uniformities in phase. It was shown that since each frequency component of the focal source signal is imaged to a planar source one focal distance behind the reflector, the entire spectrum of frequency components in the short pulse is imaged to that plane, and the parabola can be replaced by that planar source (with appropriate amplitude weighting) illuminating an aperture in an conducting screen.

For defocused sources, which give scanned beams, the reflection properties of the parabola are quite frequency dependent. In this case, there are geometric path length errors for every incident ray. For each frequency component the physical error corresponds to a different electrical phase error. The short time pulse will reflect very differently from a constant frequency source, and the transient antenna response will be entirely different from a tilted plane wave illuminating an equivalent aperture.

## REFERENCES

1. Hansen, R., "Short-Pulse Excitation of Reflector Antennas," *IEE Proceedings*, vol. 134, Pt. H, no. 6, December 1987, pp. 557-559.
2. Hansen, R., "Short-Pulses in Reflectors Revisited," *IEE Proceedings*, Pt. H.
3. Yee, K.S., "Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media", *IEEE Transactions on Antennas and Propagation*, vol. AP-14, 1966, pp. 302-307.
4. Taflove, A., and Umashankar, K., "The Finite-Difference Time-Domain (FDTD) Method for Electromagnetic Scattering and Interaction Problems," *Journal of Electromagnetic Waves and Applications*, vol. 1, 1987, pp. 243-267.
5. Borup, D., Sullivan, D., and Gandhi, O., "Comparison of the FFT Conjugate Gradient Method and the Finite-Difference Time-Domain Method for the 2-D Absorption Problem," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-35, 1987, pp. 383-395.
6. Luebbers, R., Hunsberger, F., and Kunz, K., "FDTD Formulation for Frequency Dependent Permittivity," *AP-S Sym. Digest* June 1989, pp. 50-53.
7. Ling, R., "Application of Computational Fluid Dynamics Methods to a Numerical Study of Electromagnetic Wave Scattering Phenomena," *Journal of Applied Physics*, vol. 64, 1988, pp. 3785-3791.
8. Rappaport, C. and McCartin, B., "FDFD Analysis of Electromagnetic Scattering in Anisotropic Media Using Unconstrained Triangular Meshes," *IEEE Transactions on Antennas and Propagation*, vol. AP-39, no. 3, March 1991, pp 345-349.
9. Rappaport, C. and Smith, E., "Anisotropic FDFD Computed on Conformal Meshes." *Fourth Biennial IEEE Conference on Electromagnetic Field Computation*, October 1990, p. BB-10.

10. Bayliss, A., Gunzburger, M., and Turkell, E., "Boundary Conditions for the Numerical Solution of Elliptic Equations in Exterior Regions," *SIAM Journal of Applied Mathematics*, vol. 42, 1982, pp. 430-451.
11. Engquist, B., and Majda, A., "Absorbing Boundary Conditions for the Numerical Simulation of Waves," *Mathematical Computation*, vol. 31, 1977, pp. 629-651.
12. Kriegsmann, G.A. and Morawetz, C.S., "Solving the Helmholtz Equation for Exterior Problems with Variable Index of Reflection", *SIAM Journal of Sciences, Statistical Computation*, vol. 1, 1980, pp. 371-385.
13. Lee, C., Shin, R., Kong, J., and McCartin, B.J., "Absorbing Boundary Conditions on Circular and Elliptic Boundaries," *Prog. in Electromag. Research Sym. Proceedings*, July 1989, pp. 317-318.
14. Rappaport, C., and Bahrmassel, L., "An Absorbing Condition Based on Anechoic Chamber Absorber," (invited paper) *1991 Progress in Electromagnetics Research Symposium Proceedings*, July 1991.

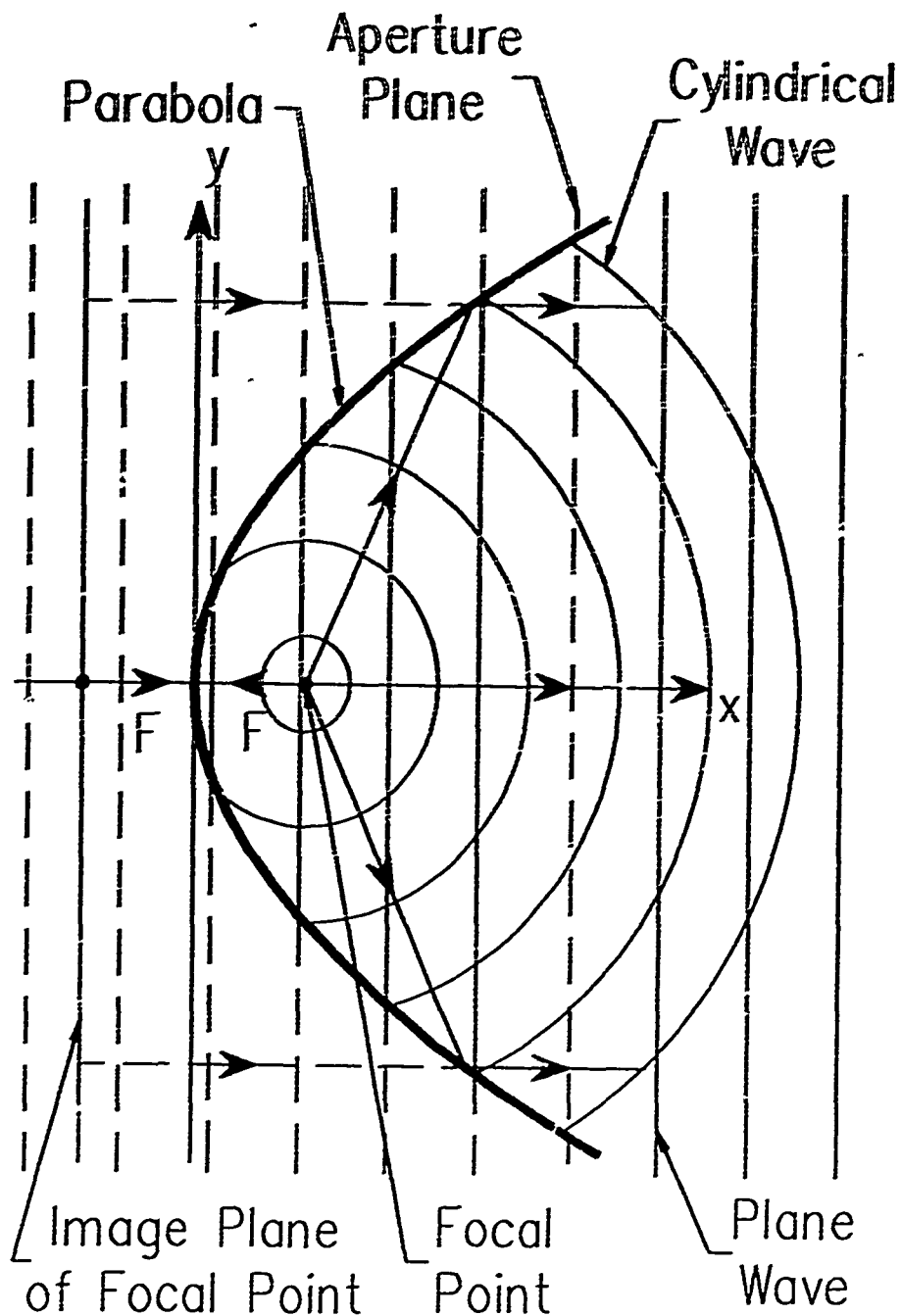
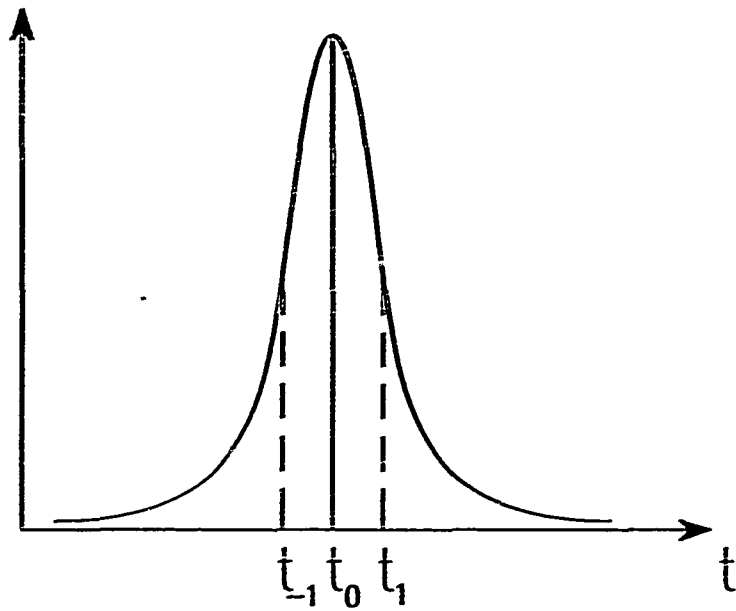
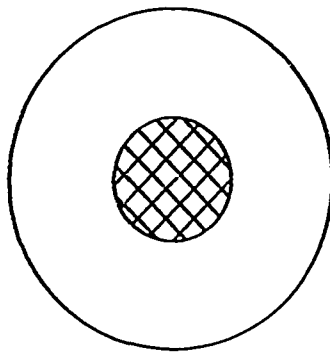


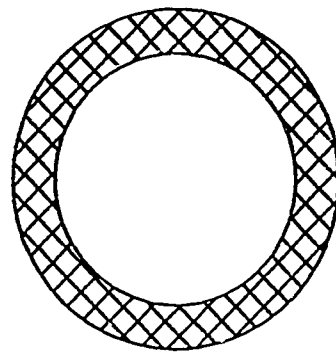
Figure 1: A short pulse in time, and the different illuminated regions of a paraboloidal reflector surface.



Short Time Pulse



$$t < t_{-1}$$



$$t > t_1$$

Illuminated Region of  
Paraboloidal Reflector

Figure 2: Ray tracing and wave reflection by a parabola.



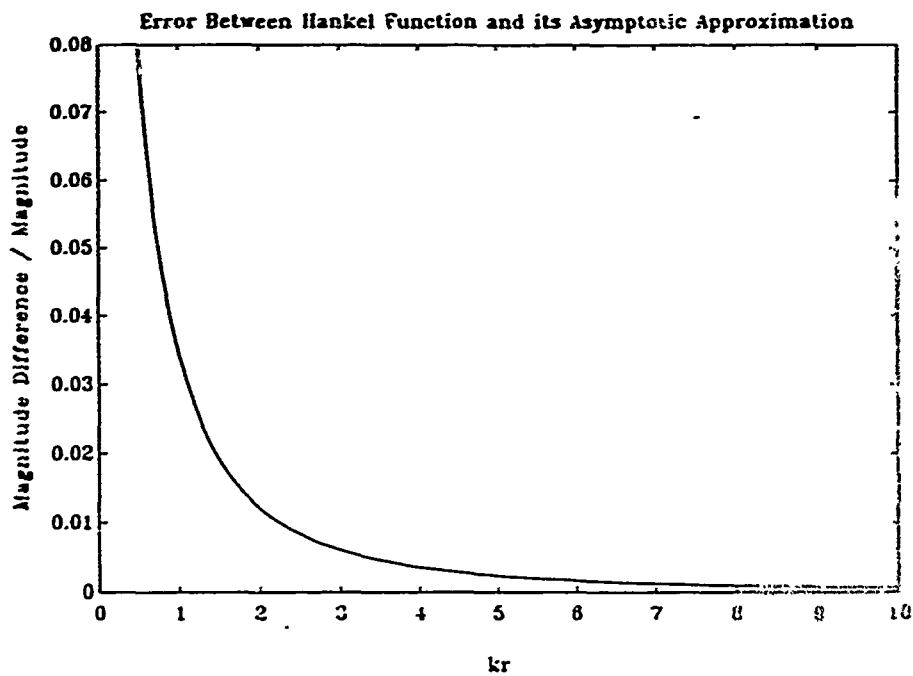
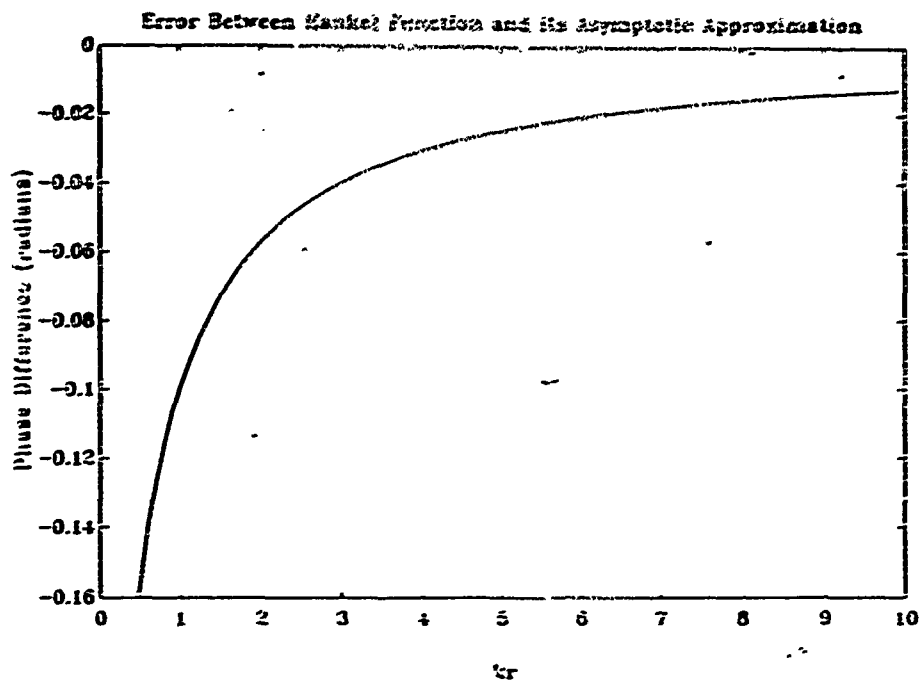
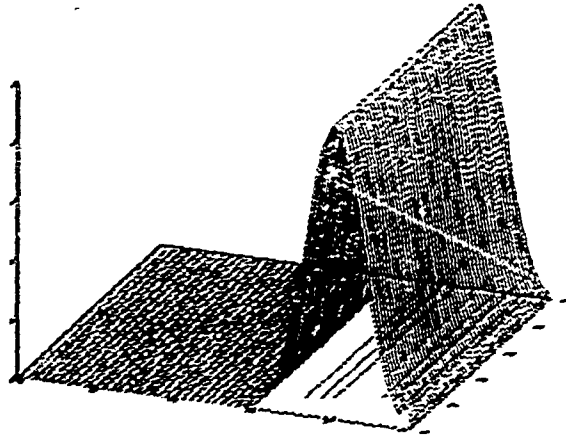
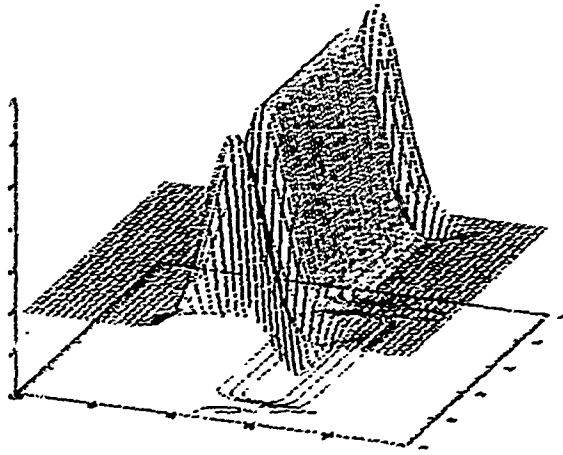


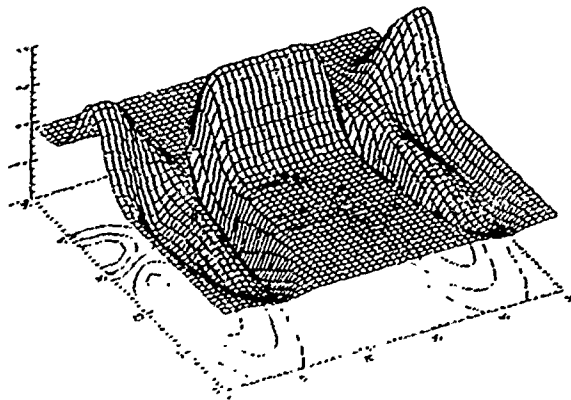
Figure 3: Comparison of the Hankel function and its asymptotic approximation.



a)

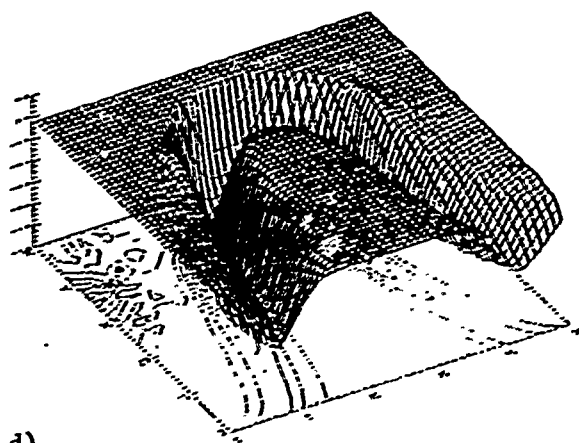


b)

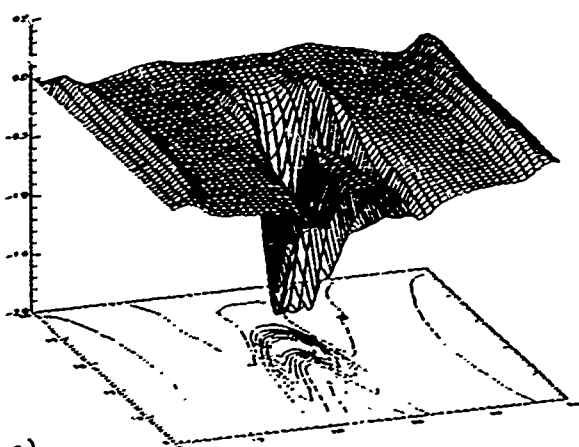


c)

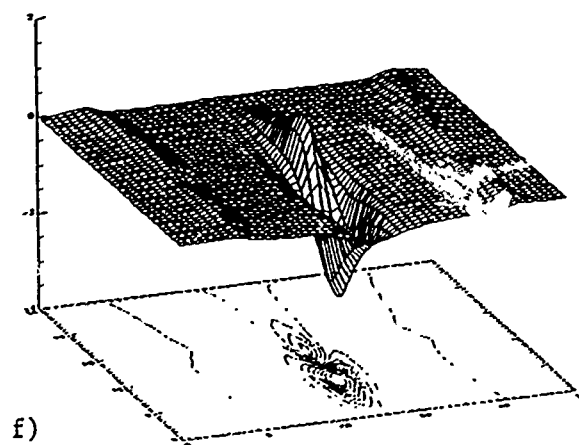
Figure 4: Uniform plane wave gaussian pulse received by a parabolic reflector on a 200 by 500 point grid at times  $t = 0, 100, 200, 300, 400, 500$ .



d)



e)



f)

Figure 4 (continued)

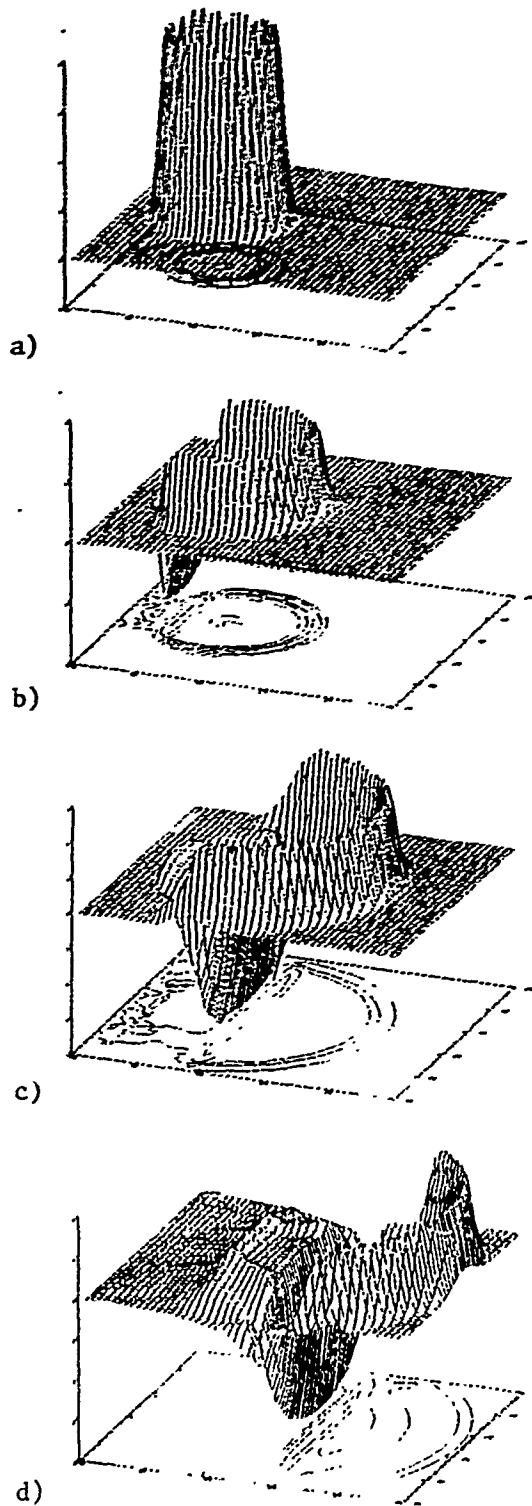


Figure 5: Cylindrical wave gaussian pulse transmitted by a parabolic reflector on a 500 by 500 point grid at times  $t = 50, 150, 350, 550$ .

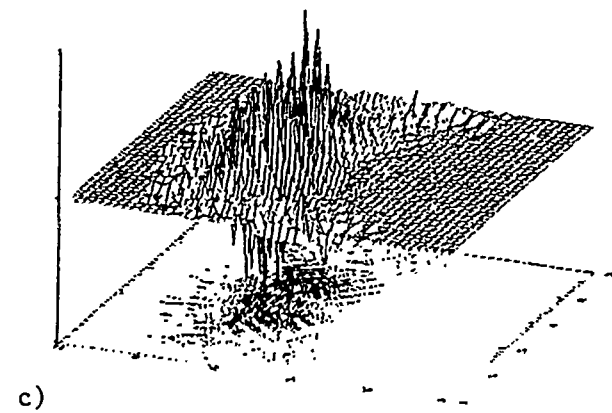
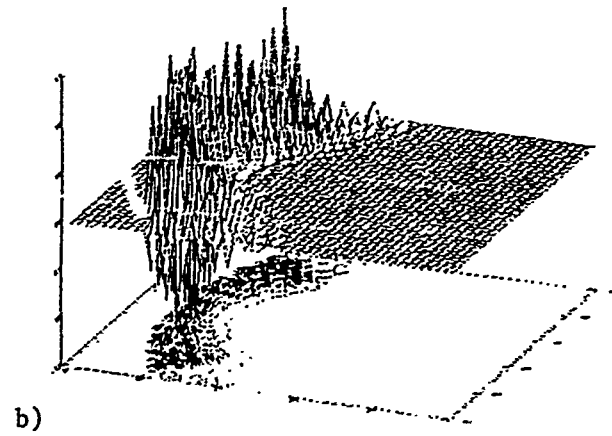
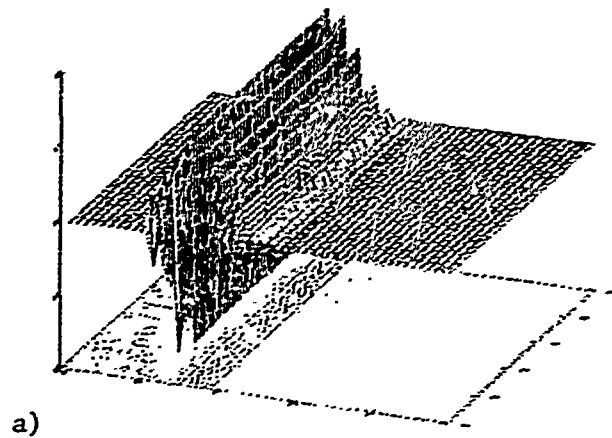


Figure 6: Modulated plane wave gaussian pulse received by a parabolic reflector on a 500 by 500 point grid at times  $t = 200, 300, 400, 480, 550, 650$ .

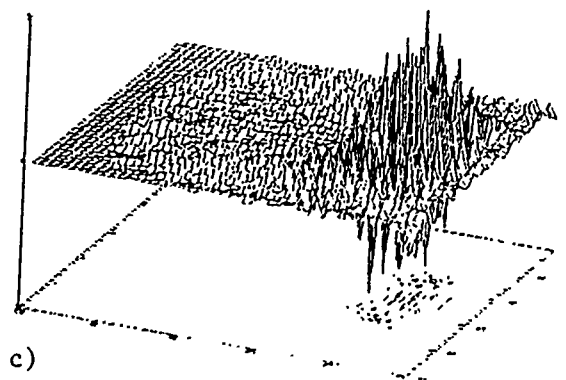
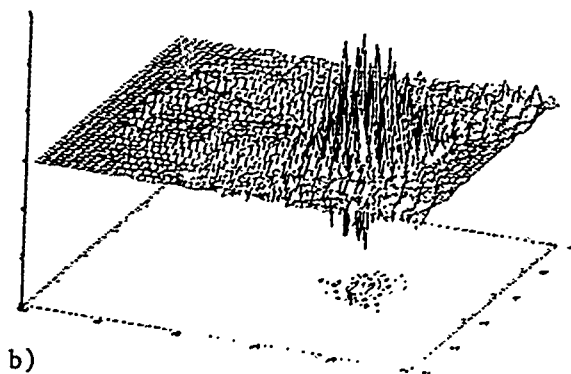
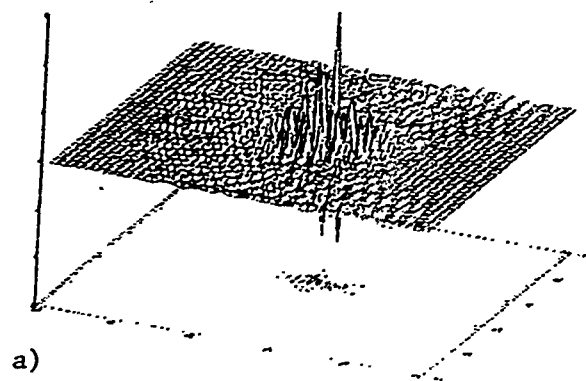


Figure 6 (continued)

# **THERMODYNAMIC AND TRANSPORT PROPERTIES OF MOLECULAR IONS**

## **FINAL REPORT**

**PREPARED BY:** MURTY A. AKUNDI, Ph.D

**ACADEMIC RANK:** PROFESSOR

**UNIVERSITY AND DEPARTMENT:** XAVIER UNIVERSITY  
PHYSICS/ENGINEERING  
NEW ORLEANS, LA 70125

**AFSOR/AEDC/CALSPAN**

**DIVISION:** COMPUTATIONAL FLUID  
DYNAMICS

**FOCAL POINT:** JAMES T. CURTIS (CALSPAN)

**DATE SUBMITTED:** AUGUST 3, 1991

**EMPLOYEE NUMBER:** 2

## ABSTRACT

Thermodynamic properties, enthalpy, free energy and heat capacity of six ionized species ( $\text{CO}^+$ ,  $\text{C}_2^-$ ,  $\text{H}_2\text{O}^+$ ,  $\text{H}_3^+$ ,  $\text{H}_3\text{O}^+$ , and  $\text{HCO}^+$ ) are calculated using spectroscopic data. These thermodynamic properties are a function of temperature and were calculated to a temperature of 20,000 K. Results are compared with the values obtained from earlier workers.

## INTRODUCTION

Currently envisaged transatmospheric and aeroassist missions have created a resurgence of interest in the aerothermic design of hypersonic vehicles. However, the velocities and altitudes at which the aircraft would operate are unfamiliar and quite different and severe. As a result, the non-equilibrium flow environment which will surround these vehicles will considerably impact the vehicle aerodynamics, thermal loads, and propulsion system efficiency. The design of future vehicles suitable to these conditions, requires careful physical modeling to simulate these phenomena.

Under hypersonic flight conditions, a vehicle traveling through the atmosphere will excite the air, which flows around the body, to very high temperatures as the kinetic energy from the vehicle is transferred. This will affect the properties of air and cause deviation from those of a thermally and calorically perfect gas. Under these conditions, the thermodynamic and transport properties of the mixture becomes a function of pressure as well as temperature. It is essential to evaluate these functions in order to calculate



the pattern of air flow about high speed vehicles, the viscous and pressure forces which result, and the heat flux which occurs between the air and the vehicle. Thermodynamic and transport properties of several species in air have been calculated by Gupta, et al., (1989) to very high temperatures.

In addition entry probe heat shields are often made of carbonaceous material such as carbon-phenolic ablator, a mechanism used for reducing the intensive radiative heating encountered during the entry into the upper atmosphere (Moss, et al. 1975, 1976). These heat shields typically contain 92% carbon, 6% oxygen, and 2% hydrogen by mass (Moss, et al. 1976). This ablation injects gaseous carbonaceous materials which drastically alter the atmospheric composition near the entry vehicle, which in turn changes the thermodynamic and transport properties. Thus, enthalpy, free energy, heating, diffusion, and viscous flow rates are changed.

Biolsi (1978), Rainwater, et al. (1981), and Biolsi, et al. (1981) investigated the transport properties of monoatomic carbon as well as a mixture of the ablation products C, C<sub>2</sub>, and C<sub>3</sub>. In this report we present the thermodynamic properties of CO<sup>+</sup>, C<sub>2</sub><sup>+</sup>, H<sub>2</sub>O<sup>+</sup>, H<sub>3</sub><sup>+</sup>, H<sub>3</sub>O<sup>+</sup>, and HCO<sup>+</sup> that are computed using a simple computer program developed at this laboratory. A general discussion on the method to calculate the diffusion, viscosity, and thermal conductivity from collision integrals using H-H potentials is provided. Due to the non-availability of a computer program, transport properties of these molecules could not be calculated. The future plan is to develop a suitable computer program to evaluate the collision integrals and calculate the transport properties.

## PROBLEM DISCUSSION

Entry probe heat shields, at high temperatures, inject ablative gaseous products such as ionized carbon monoxide, carbon, water, and several other combinations. Data on the thermodynamic and transport properties of these species at high temperatures ( $>6,000$  K) is not available in the literature. Determination of these properties is vital to understand the air flow patterns which in turn aid in the thermodynamic design of the hypersonic vehicles.

### Thermodynamic Properties:

The equilibrium thermodynamic properties of gas can be calculated with certainty provided the energy levels of the gas particles and the degeneracy of these levels are known. For most of the diatomic and for some of the polyatomic gaseous species, this information can be deduced from spectroscopic data. All of the thermodynamic properties, heat capacity ( $C_p$ ), enthalpy ( $H^\circ$ ), free energy ( $G^\circ$ ), and entropy ( $S^\circ$ ) of the species can be determined from a knowledge of their partition functions ( $Q$ ). Consequently, the first step is to determine the partition functions of these species, which is defined as (Herzberg, 1945, McBride, et al., 1963)

$$Q = \sum g_i \exp - \left[ E_i / K_B T \right] \quad (1)$$

where  $E_i$  is the energy of the  $i^{\text{th}}$  state of the particle and  $g_i$  is the degeneracy. The energy may be due to the translational, vibrational, or rotational motion of the particle or due to the motion of electrons within the particle. Then the partition function given in Eq. (1) can be expressed as a product

$$Q = Q_t Q_e Q_v Q_r Q_p Q_w Q_c \quad (2)$$

where  $Q_t$ ,  $Q_e$ ,  $Q_v$ ,  $Q_r$ ,  $Q_p$ ,  $Q_w$ , and  $Q_c$  are the translational, electronic, vibrational, rotational, rotational stretching, Fermi resonance and both anharmonicity and vibration rotation interaction, respectively. These individual partition functions can be expressed as functions of temperature as

$$Q_t = \left[ \frac{2\pi M K_B T}{h^2} \right]^{3/2} \frac{RT}{P} \quad (3)$$

For diatomic and linear polyatomic molecules

$$Q_r = \left[ \frac{8\pi^2 I K_B T}{\sigma h^2} \right] \quad (4)$$

For non-linear polyatomic molecules

$$Q_r = \sigma^{-1} \pi^{1/2} \frac{(8\pi^2 I_A K_B T)^{1/2}}{h^2} \frac{(8\pi^2 I_B K_B T)^{1/2}}{h^2} \frac{(8\pi^2 I_C K_B T)^{1/2}}{h^2} \quad (5)$$

For diatomic

$$Q_v = \left( 1 - e^{-\frac{h\nu c}{K_B T}} \right)^{-1} \quad (6)$$

For polyatomic

$$Q_v = \left( 1 - e^{-\frac{h\nu_i c}{K_B T}} \right)^{-d_i} \quad (7)$$

$$Q_c = \sum_{n=0}^{\infty} \frac{g_n}{n!} e^{-E_n / k_B T} \quad (8)$$

**Polyatomic**

$$Q_c = 2\pi / \beta \sum_i d_i n_i^1 \phi_i + \frac{1}{2} \sum_i d_i (d_i + 1) X_{ii}^{-1} \phi_i = \sum_{i < j} d_i d_j X_{ij}^{-1} \phi_i \phi_j \quad (9)$$

The properties ( $C_p$ ,  $H^\circ$ , and  $G^\circ$ ) can then be expressed as a function of  $Q$ , that is

$$C_p / R = \frac{T^2 d^2 (\ln Q)}{dT^2} + 2T \frac{d}{dT} (\ln Q) + \frac{5}{2} \quad (10)$$

$$\frac{H_T^\circ - H_o^\circ}{RT} = T \frac{d}{dT} (\ln Q) + \frac{5}{2} \quad (11)$$

$$\frac{-G(T) - H_o^\circ}{RT} = \ln Q + \frac{3}{2} \ln M + \frac{5}{2} \ln T + S_c \quad (12)$$

where

$$S_c = \ln \left[ K_B \left( \frac{2\pi K_B}{N_O h^2} \right)^{3/2} \right] \quad (13)$$

As can be seen from Eqs. (10) to (13), the translational motion of the particle contributes only to the free energy of the particle. In the calculations made in this report, the contributions due to rotational stretching and Fermi resonance are not considered. Details of the calculations and the values obtained for these properties are presented in the results section.

## Transport Properties:

Transport properties of gases may be evaluated according to the well known Chapman-Enskog solution of the Boltzman equation (Herschfelder, et al., 1954) if the intermolecular potential is known. These intermolecular potentials can be easily evaluated for diatomic molecules but it is more involved for polyatomic molecules. Within the first Enskog approximation, the self diffusion coefficient of a gas is inversely proportional to  $\Omega(1, 1)^*$  integral and the shear viscosity and thermal conductivity are inversely proportional to  $\Omega(2, 2)^*$  integral. The general form of the reduced collision integral  $\Omega(\ell, s)^*$  is represented as

$$\Omega(\ell, s)^* = C_{\ell, s} \int E^{s-1} e^{-E/K_B T} dE \int_0^\infty b db (1 - \cos^\ell X) \quad (14)$$

where  $C_{\ell, s}$  is a constant ( $C_{1,1} = 1$ ,  $C_{2,2} = \frac{1}{2}$ ),  $K_B$  is Boltzman's constant,  $T$  the temperature,  $E$  and  $b$  the relative total energy and impact parameter and  $X$  the scattering angle given by

$$X = \pi - 2b \int_{r_m}^\infty \frac{dr}{r^2 \left[ 1 - \frac{Q_{eff}(r)}{E} \right]^{\frac{1}{2}}} \quad (15)$$

where

$$Q_{eff} = \phi(r) + (Eb^2/r^2)$$

and

$$\phi(r) = \exp[-2a(r/d - 1)] - 2 \exp[-a(r/d - 1)]$$

$$\div \beta(r/d - 1)^3 [1 \div r/d - 1] \exp[-2a(r/d - 1)] \quad (16)$$

where

$$a_0 = \omega_c^2 / 4Be, a_1 = -1 - a \omega_c / 6B_c^2$$

$$a_2 = 5/4 a_1^2 - 2 \omega_c X / 3B_c$$

$$a = \omega_c / 2 (BeE)^{1/2}, \beta = a^3 [1 \div a, (E^1/a\omega)^{1/2}]$$

$$\gamma = a \{2 - a^3 \beta^{-1} (7/12 - E^1 a^2 / a\omega)\}$$

After obtaining the numerical values of  $\Omega(1,1)$  and  $\Omega(2,2)$  by evaluating the collision integrals, the transport properties can be calculated using the following equations

$$\text{Viscosity: } \eta = \frac{5}{16} \sqrt{\frac{M k_B T}{\pi}} \cdot \frac{1}{\sigma^2 \Omega(2,2)} \quad (17)$$

$$\text{Thermal Conductivity: } \lambda = \frac{25}{32} \sqrt{\frac{k_B T}{\pi M}} \cdot \frac{c_v}{\sigma^2 \Omega(2,2)} \quad (18)$$

$$\text{Diffusion: } D = \frac{3}{8} \sqrt{\frac{k_B T}{\pi M}} \cdot \frac{1}{8 \sigma^2 \Omega(1,1)} \quad (19)$$

The problem at hand is to develop numerical integral methods to evaluate the collision integrals. Recently Rainwater, etc. (1982) have

developed a computer program to evaluate these integrals. At present we are in the process of obtaining this program and modifying it for our conditions. The symbols and their meaning of the various parameters used in the equations are presented in the Appendix.

## RESULTS

Literature survey has been made and relevant molecular parameters are collected for the six species to compute the thermodynamic properties. In the case of molecules for which the needed data is not available, the necessary parameters (dissociation energies, force constants, and excited state electronic energies) are computed using appropriate equations given by Hertzberg. Heat capacity, enthalpy, and free energy are calculated using the computer program developed earlier in this laboratory by Dr. James T. Curtis. Table I presents the molecular constants that are used in calculating the partition function. Tables II to VII present the thermodynamic properties of the six species to a temperature of 20,000 K. These data have been plotted and compared with the data in JANNAF, and shown in Figs. 1 to 3. Though thermodynamic data for some of these molecules ( $C_2$ ,  $H_3^+$ , and  $HCO^+$ ) are calculated by earlier workers (JANNAF, 1974, McBride, 1967), they are only computed up to 6,000 K. In the case of  $HCO^+$  and  $H_3O^+$ , the vibrational frequencies are calculated using the spectroscopic data of isoelectronic species  $NH_3$  and  $HCN$ . In the present report, spectroscopic data collected is more recent and accurate and is based on experimental work. In addition, since the thermodynamic properties are calculated to high temperatures, this data will be more useful to understand the non-equilibrium air flow conditions and to

a precise aerothermic design of the space vehicles. In  $\text{H}_3\text{O}^+$  and  $\text{HCO}^+$ , only ground state electron energy is used in evaluating the electronic partition function, since the excited state energy values are not available at the present time. For  $\text{CO}^+$  and  $\text{C}_2^+$ , all partition function contributions (except  $Q_p$  and  $Q_w$ ) are used to evaluate the thermodynamic properties. On the other hand, in the case of polyatomic molecules  $\text{H}_2\text{O}^+$ ,  $\text{HCO}^+$ ,  $\text{H}_3^+$ , and  $\text{H}_3\text{O}^+$  only contributions of electronic, rotational, and vibrational partitions function are considered, since complete data on anharmonic constants and rotation vibration interaction terms on the species is not available at the present time.

## CONCLUSIONS

Heat capacity, enthalpy, and free energy of  $\text{CO}^+$ ,  $\text{C}_2^+$ ,  $\text{H}_2\text{O}^+$ ,  $\text{H}_3^+$ , and  $\text{H}_3\text{O}^+$  have been calculated up to a temperature of 20,000 K. Vibration rotation interaction and anharmonicity influence on thermodynamic properties is significant at high temperatures. In the case of  $\text{CO}^+$  and  $\text{C}_2^+$ , this effect is included in our calculations. Unfortunately, we could not study this effect in the remaining molecules due to insufficient data. There is a need to improve this data and our future work will focus on improving the current computer program to include the affect of anharmonicity and vibration rotation interaction. Transport properties of these six species could not be computed due to the non-availability of a suitable computer program and due to the short time I have spent here. The current plan is to keep in contact with the CFD group and continue the work during the academic year.



## **ACKNOWLEDGEMENTS**

I wish to take this opportunity to thank AFOSR/RDL summer faculty program for giving me an opportunity to work at AEDC. My sincere thanks to Dr. James T. Curtis for introducing me to this fascinating new field of research and for his valuable guidance throughout the project period. It will be unfair on my part if I do not express my appreciation to Mr. M. Aboulmouna for his encouragement and constant help with the computation of thermodynamic properties. Last but not least, I wish to thank all Calspan personnel in the Computational Fluid Dynamics Division who made my stay here very pleasant and rewarding.

## REFERENCES

1. Amano, T. A., *J. Chem. Phys.*, 79, 3595 (1983).
2. Begemann, M. H., and R. J. Saykally, *J. Chem. Phys.*, 82, 3570 (1985).
3. Biolsi, L., *J. Geo. Phys. Res.*, 83, 2476 (1978).
4. Biolsi, L., J. Fenton, and B. Owenson, *Progress in Astronaut. and Aeronaut.*, 82, 17 (1981).
5. Davies, P. B., and W. J. Rothwell, *J. Chem. Phys.*, 81, 5239 (1984).
6. Dinelli, B. M., M. W. Crofton, and T. Oka, *J. Mol. Spectrosc.*, 127, 1, (1988).
7. Fortune, P. J., B. J. Rosenberg, and A. C. Wahi, *J. Chem. Phys.*, 65, 2201 (1974).
8. Foster, S. C., A. R. W. McKellar, and T. J. Sears, *J. Chem. Phys.*, 81, 578 (1984).
9. Gupta, R. N., J. M. Yos, and R. A. Thompson, NASA TM 101528 (1989).
10. Henderson, J. R., and J. Tennyson, *Chem. Phys. Lett.*, 173, 133 (1990).
11. Herschfelder, J. O., C. F. Curtis, and R. B. Bird, "Molecular Theory of Gases and Liquids", John Willey and Sons, New York, (1954), Chs. 7 and 8.
12. Hergberg, G., "Infrared and Raman Spectra", *Molecular Spectra and Molecular Structure II*, (1945).
13. Huber, K. P., and G. Herzberg, *Molecular Spectra and Molecular Structure IV*, (1979).
14. Kauppi, E., and L. Halonen, *Chem. Phys. Lett.*, 169, 393 (1990).
15. JANNAF Tables: "Thermo Chemical Data" Dow Chemical Co., (1985).
16. Kramer, W. P., and G. H. F. Dierckson, *Astro. Phys. J.*, 205, L97 (1976).
17. Leclerc, J. C., J. Horsley, and J. C. Lorquet, *Chem. Phys.*, 4, 337 (1974).

18. Liu, D. J., T. Oka, and T. J. Sears, *J. Chem. Phys.*, 84, 1312 (1986).
19. Liu, D. J., T. J. Lee, T. Oka, *J. Mol. Spectrosc.*, 128, 236 (1988).
20. McBride, B. J., and S. Gordon, NASA TN D-4097 (1967).
21. McBride, B. J., S. Heime1, J. G. Ehlers, and S. Gordon, NASA SP-3001 (1963).
22. Miller, S., and J. Tennyson, *J. Mol. Spectrosc.*, 126, 183 (1987).
23. Misra, P., D. W. Ferguson, and K. Naraharirao, *J. Mol. Spectrosc.*, 125, 54 (1987).
24. Moss, J. N., E. C. Anderson, and C. W. Boltz, Jr., Paper presented at the 10th Thermophysics Conference, Amer. Inst. of Aeronaut. and Astronaut., Denver, CO (1975).
25. Moss, J. N., E. C. Anderson, and C. W. Boltz, Jr., Paper presented at the 11th Thermophysics Conference, Amer. Inst. of Aeronaut. and Astronaut., San Diego, CA (1976).
26. Rainwater, J. C., P. M. Holland, and L. Biolsi, *Prog. in Astronaut. and Aeronaut.*, 82, 3 (1981).
27. Rainwater, J. C., J. F. Ely, and H. J. M. Hanley, (Private Communication).
28. Rehfuss, B. D., D. J. Liu, B. N. Dinelli, M. F. Jagod, C. W. Ho, M. W. Crofton, and T. Oka, *J. Chem. Phys.*, 89, 129 (1988).

TABLE IA  
MOLECULAR CONSTANTS (DIATOMIC GASES)

Species	Mol. wt	Sym. (o)	Elec State	Statis $\omega_t$ (g)	To (K)	$\omega_e$ cm <sup>-1</sup>	$\omega_{exe}$ cm <sup>-1</sup>	B <sub>e</sub>	$\alpha_e$	D <sub>e</sub> x10 <sup>-6</sup> (cm <sup>-1</sup> )	D <sub>0</sub> <sup>a</sup> (cm <sup>-1</sup> )	Ref
CO <sup>+</sup>	28.0104	1	2 $\Sigma^+$	2	0	2214.27	15.20	1.976959	0.018975	6.327	80641	23
			2 $\Pi$	2	29830	1562.06	13.532	1.5874	0.01940	2.8	..	
C <sub>2</sub>	24.0225	2	2 $\Sigma^+$	2	0	1781.189	11.672	1.7467	0.01651	6.69	69306.7	13,28
			2 $\Pi$	4	5695.9	1666.4	10.8	1.6431	0.01601	7.27	..	
			2 $\Sigma^-$	2	26594.5	1968.7	14.43	1.8774	0.1776	6.84	..	

TABLE IB  
(POLYATOMIC MOLECULES)

Species	Mol. wt	Sym. (o)	Elec State	Statis $\omega_t$	To (K)	$\omega_1$ (cm <sup>-1</sup> )	$\nu_2$ (cm <sup>-1</sup> )	$\nu_3$ (cm <sup>-1</sup> )	I <sub>A</sub> (10 <sup>-39</sup> ) g cm <sup>2</sup>	I <sub>B</sub> (10 <sup>-39</sup> ) g cm <sup>2</sup>	I <sub>C</sub> (10 <sup>-39</sup> ) g cm <sup>2</sup>	D <sub>0</sub> <sup>a</sup> (cm <sup>-1</sup> )	Ref
HCO <sup>+</sup>	29.178	1	1 $\Sigma^+$	1	0	3058.73	829.03	2183.95	1.882	1.876	1.894	34048	1,5,8,19
			1 $\Sigma^-$	1									
H <sub>2</sub> O <sup>+</sup>	18.015	2	2B <sub>1</sub>	2	0	3159	1494	3267	0.0964	0.2253	0.3305	31224	6,7,14,17
			2 $\Sigma^+$	2	10,550					..	..	..	
			2 $\Pi_g$	2	26,354		..	..			..	..	
H <sub>3</sub> <sup>+</sup>	3.024	2,6 <sup>+</sup>	1 $\Sigma^+$	1	0	3178.35 <sup>b</sup>	2521.28 <sup>c</sup>	4363.50 <sup>d</sup>	0.0643	0.0643	0.1359	37047	110,22
			1 $\Sigma^-$	1	163/b					..	..	..	
H <sub>3</sub> O <sup>+</sup>	19.023	3	1 $\Sigma^+$	1	0	3398	954.96	3504	0.2488	0.2618	0.2540	38163	2,18

a = Symmetry no.2 is used in these calculations  
b = listed as A  
c = listed as E  
d = Listed as ZPE

$C_p$ , J/mole K

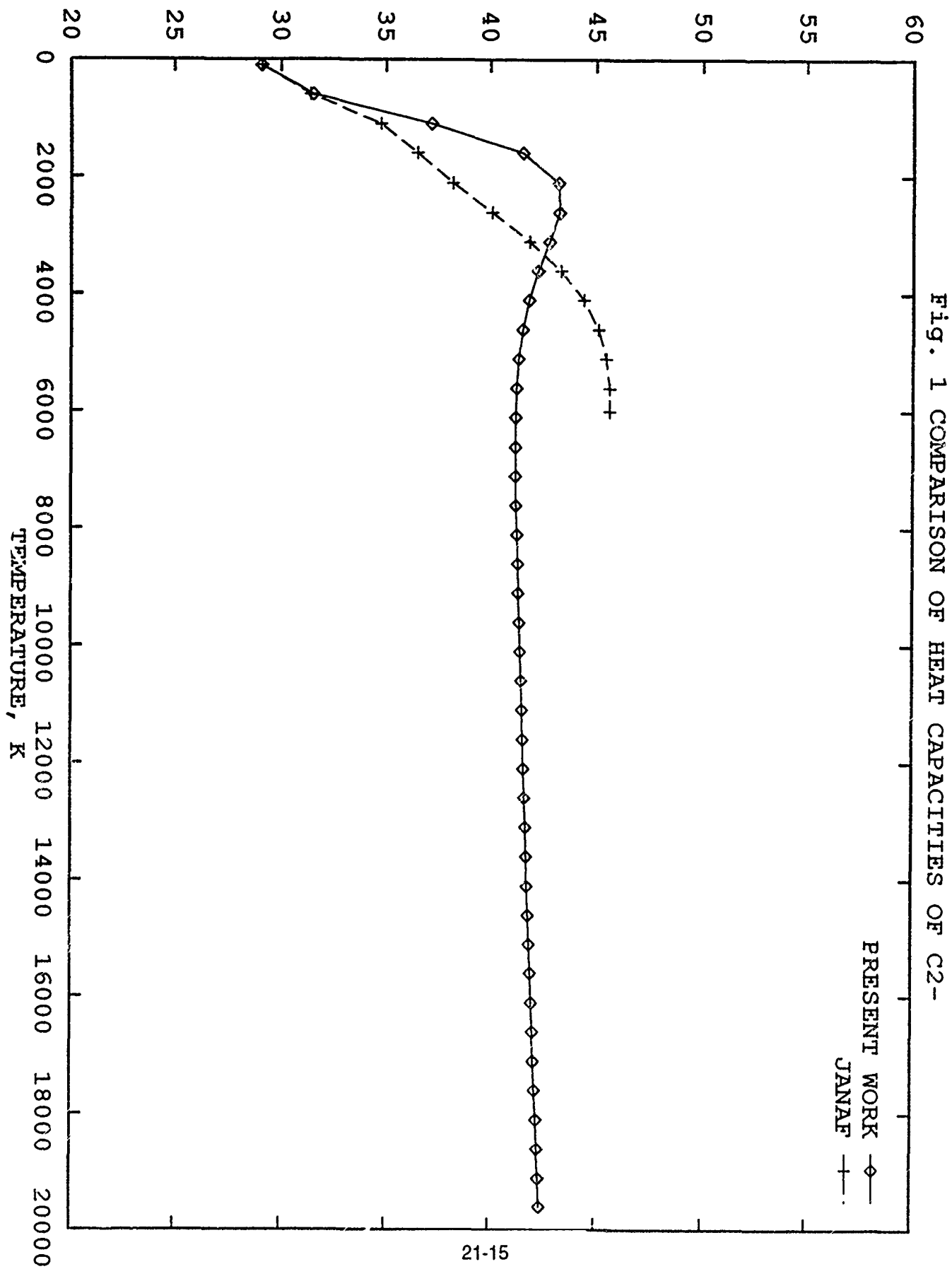


Fig. 2 COMPARISON OF HEAT CAPACITIES OF H3O+

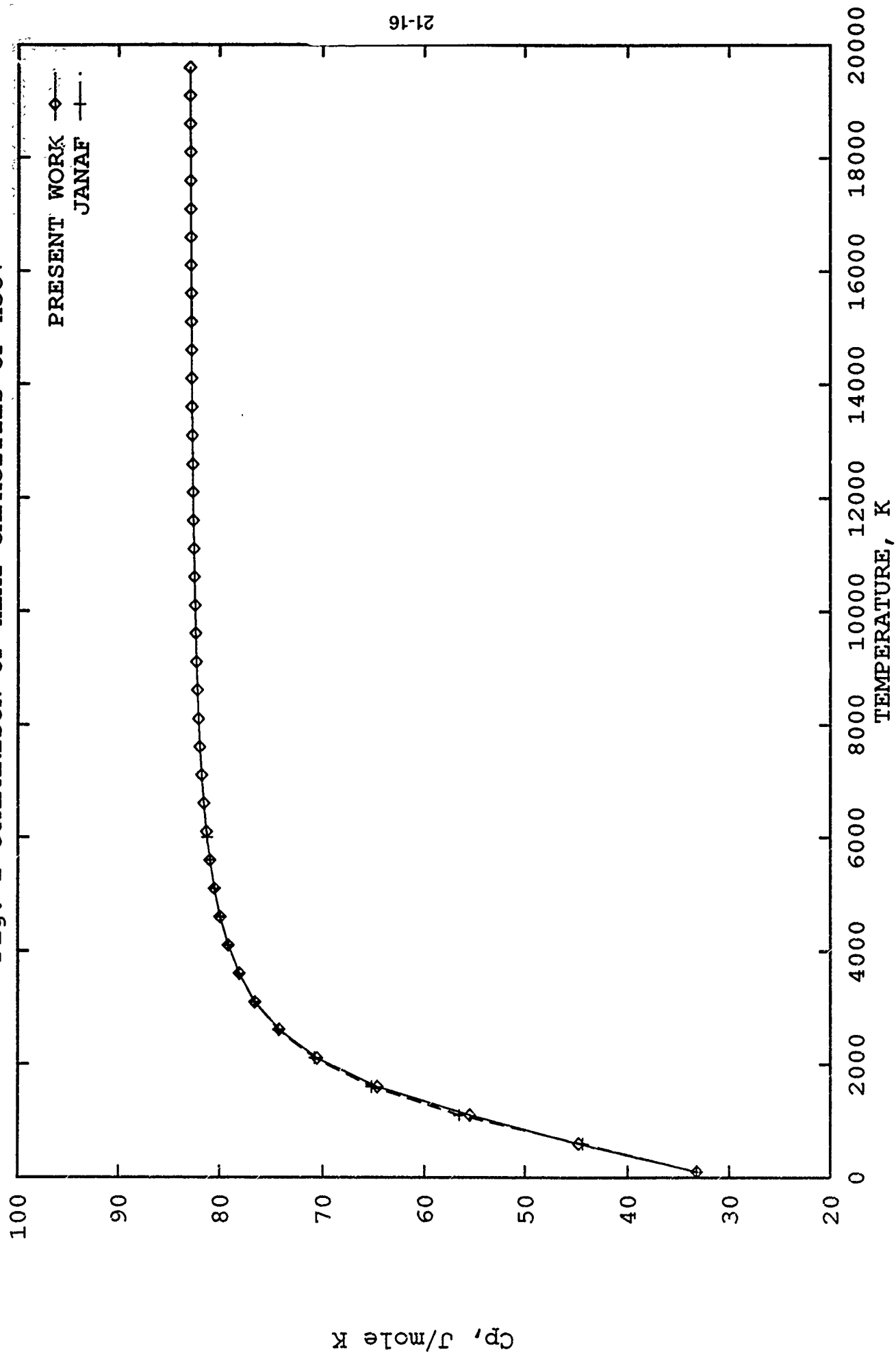
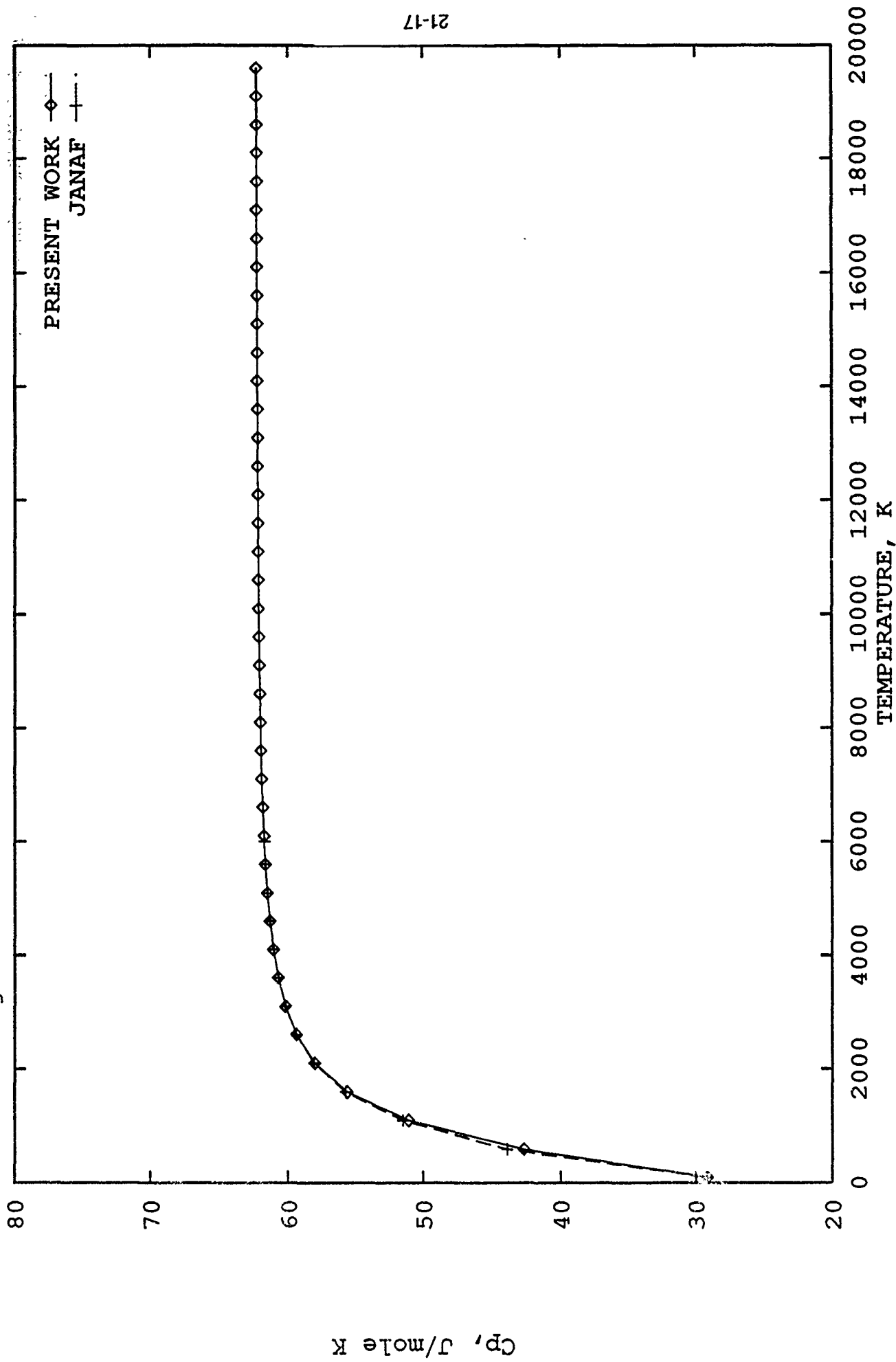


Fig. 3 COMPARISON OF SPECIFIC HEAT CAPACITIES OF  $\text{HCO}^+$ 

## **APPENDIX A**



## APPENDIX A SYMBOLS

$A_e, B_e, C_e$	Rotational constants corresponding to equilibrium separation of atoms, $\text{cm}^{-1}$ .
$A_0, B_0, C_0$	Rotational constants for lowest vibrational states, $\text{cm}^{-1}$ .
$C_p^\circ$	Heat capacity at constant pressure of standard state cal/mole. K.
$c$	Velocity of light, $2.997925 \times 10^{10} \text{ cm}$ .
$C_2$	Second radiation constant, $hc/K$ , 1.43880 (cm) (deg).
$D_e, D_v$	Spectroscopic constant, for rotational stretching, $\text{cm}^{-1}$ .
$D_0, D_{00}$	Rotational stretching constant for lowest vibrational state, $\text{cm}^{-1}$ .
$D_0^\circ$	Dissociation energy measured from the lowest vibrational energy level ( $v=0$ ).
$d_i$	Degeneracy associated with $V_i$ .
$F_T^\circ$	Gibbs free energy for standard state, cal/mole.
$F_T^\circ - H_0^\circ$	Sensible energy for standard state, cal/mole.
$g_e$	Electronic statistical weight.
$g_{ii}$	Anharmonicity constant for doubly degenerate vibrations in linear molecules, $\text{cm}^{-1}$ .
$H_T^\circ$	Sum of sensible enthalpy and chemical energy at 0 K for standard state, cal/mole.
$H_0^\circ$	Chemical energy at 0 K for standard state, cal/mole.
$H_T^\circ - H_0^\circ$	Sensible enthalpy for standard state, cal/mole.
$\Delta H_T^\circ$	Enthalpy change for formation of substance for elements in atomic gas state, cal/mole.
$h$	Planck's constant, $6.6261 \times 10^{-27}$ (erg) (sec).
$I_A, I_B, I_C$	Principal moments of inertia, (g) ( $\text{cm}^2$ ) (mole).
$J$	Rotational quantum number.
$k_B$	Boltzmann constant, $1.38044 \times 10^{-16}$ erg/deg.
$M$	Molecular weight.
$n$	Number of unique frequencies.
$p$	Partial pressure, atm.
$Q$	Internal partition function.
$Q_a$	Anharmonic correction factor for partition function.
$Q_e$	Electronic partition function.
$Q^m$	Internal partition function for $m^{\text{th}}$ electronic state.
$Q_R$	Classic rotational partition function.
$Q_{RV}$	Vibration-rotation interaction correction for partition function.
$Q_V$	Harmonic-oscillator partition function due vibrational motion.
$Q_\omega$	Rotational-stretching correction factor to partition function.
$R$	Universal gas constant, 1.9876 cal/mole K.
$r$	Internuclear distance, $\text{\AA}$ .
$r_e$	Equilibrium internuclear distance, $\text{\AA}$ .
$T$	Temperature K.
$T_0$	Electronic excitation energy between lowest vibrational states ( $V=0$ ) of ground and excited state, $\text{cm}^{-1}$ .

$u_i$	Vibrational quantum number.
$\alpha_e, \alpha_j$	Vibration rotation interaction constants for diatomic molecules, $\text{cm}^{-1}$ .
$\alpha_i A, \alpha_i B, \alpha_i C$	Vibration rotation interaction constants for polyatomic molecules, $\text{cm}^{-1}$ .
$v_i$	Observed fundamental free energy, $\text{cm}^{-1}$ .
$p$	Rotational stretching spectroscopic constant, $\text{K}^{-1}$ .
$\sigma$	Symmetry number.
$\omega_e$	Zero-order vibrational frequency for diatomic molecules, $\text{cm}^{-1}$ .
$\omega_e X_e, \omega_e Y_e$	Anharmonicity constants for diatomic molecules, $\text{cm}^{-1}$ .

## **APPENDIX B**

Thermodynamic Properties of HCO<sup>+</sup>  
Table II

T ( K )	H-HR ( j/mol )	-(G-HR)/T ( j/mol )	CP (j/mol)
0.100000E+03	-0.633885E+04	0.245910E+04	0.291158E+02
0.600000E+03	0.115211E+05	0.243667E+04	0.426881E+02
0.110000E+04	0.351483E+05	0.245230E+04	0.510540E+02
0.160000E+04	0.619353E+05	0.246557E+04	0.556035E+02
0.210000E+04	0.903996E+05	0.247669E+04	0.580111E+02
0.260000E+04	0.119774E+06	0.248622E+04	0.593672E+02
0.310000E+04	0.149679E+06	0.249452E+04	0.601885E+02
0.360000E+04	0.179915E+06	0.250187E+04	0.607179E+02
0.410000E+04	0.210369E+06	0.250845E+04	0.610772E+02
0.460000E+04	0.240974E+06	0.251442E+04	0.613312E+02
0.510000E+04	0.271689E+06	0.251987E+04	0.615172E+02
0.560000E+04	0.302484E+06	0.252489E+04	0.616572E+02
0.610000E+04	0.333341E+06	0.252954E+04	0.617651E+02
0.660000E+04	0.364245E+06	0.253386E+04	0.618500E+02
0.710000E+04	0.395188E+06	0.253791E+04	0.619179E+02
0.760000E+04	0.426161E+06	0.254171E+04	0.619731E+02
0.810000E+04	0.457159E+06	0.254530E+04	0.620186E+02
0.860000E+04	0.488178E+06	0.254869E+04	0.620565E+02
0.910000E+04	0.519215E+06	0.255190E+04	0.620884E+02
0.960000E+04	0.550266E+06	0.255496E+04	0.621155E+02
0.101000E+05	0.581330E+06	0.255788E+04	0.621387E+02
0.106000E+05	0.612404E+06	0.256067E+04	0.621587E+02
0.111000E+05	0.643488E+06	0.256333E+04	0.621761E+02
0.116000E+05	0.674580E+06	0.256589E+04	0.621913E+02
0.121000E+05	0.705679E+06	0.256835E+04	0.622047E+02
0.126000E+05	0.736784E+06	0.257071E+04	0.622165E+02
0.131000E+05	0.767895E+06	0.257299E+04	0.622271E+02
0.136000E+05	0.799011E+06	0.257519E+04	0.622364E+02
0.141000E+05	0.830132E+06	0.257731E+04	0.622448E+02
0.146000E+05	0.861256E+06	0.257937E+04	0.622524E+02
0.151000E+05	0.892384E+06	0.258136E+04	0.622592E+02
0.156000E+05	0.923515E+06	0.258328E+04	0.622654E+02
0.161000E+05	0.954649E+06	0.258515E+04	0.622710E+02
0.166000E+05	0.985786E+06	0.258697E+04	0.622762E+02
0.171000E+05	0.101693E+07	0.258873E+04	0.622808E+02
0.176000E+05	0.104807E+07	0.259045E+04	0.622851E+02
0.181000E+05	0.107921E+07	0.259211E+04	0.622891E+02
0.186000E+05	0.111036E+07	0.259374E+04	0.622927E+02
0.191000E+05	0.114150E+07	0.259532E+04	0.622960E+02
0.196000E+05	0.117265E+07	0.259687E+04	0.622991E+02

Thermodynamic Properties of H3O+  
Table III

T ( K )	H-HR ( J/mol)	-(G-HR) /T ( J/mole)	CP (J/mole)
0.100000E+03	-0.683424E+04	0.212649E+03	0.332608E+02
0.600000E+03	0.124112E+05	0.189355E+03	0.447733E+02
0.110000E+04	0.374413E+05	0.206020E+03	0.554770E+02
0.160000E+04	0.676166E+05	0.220302E+03	0.646675E+02
0.210000E+04	0.101552E+06	0.232624E+03	0.705981E+02
0.260000E+04	0.137839E+06	0.243453E+03	0.742606E+02
0.310000E+04	0.175593E+06	0.253100E+03	0.765901E+02
0.360000E+04	0.214299E+06	0.261788E+03	0.781363E+02
0.410000E+04	0.253650E+06	0.269684E+03	0.792050E+02
0.460000E+04	0.293454E+06	0.276915E+03	0.799705E+02
0.510000E+04	0.333588E+06	0.283582E+03	0.805358E+02
0.560000E+04	0.373967E+06	0.289764E+03	0.809641E+02
0.610000E+04	0.414536E+06	0.295526E+03	0.812960E+02
0.660000E+04	0.455252E+06	0.300920E+03	0.815580E+02
0.710000E+04	0.496085E+06	0.305990E+03	0.817685E+02
0.760000E+04	0.537014E+06	0.310772E+03	0.819399E+02
0.810000E+04	0.578020E+06	0.315297E+03	0.820813E+02
0.860000E+04	0.619091E+06	0.319590E+03	0.821993E+02
0.910000E+04	0.660216E+06	0.323675E+03	0.822987E+02
0.960000E+04	0.701387E+06	0.327569E+03	0.823833E+02
0.101000E+05	0.742598E+06	0.331290E+03	0.824559E+02
0.106000E+05	0.783842E+06	0.334853E+03	0.825185E+02
0.111000E+05	0.825115E+06	0.338271E+03	0.825730E+02
0.116000E+05	0.866414E+06	0.341554E+03	0.826207E+02
0.121000E+05	0.907735E+06	0.344713E+03	0.826627E+02
0.126000E+05	0.949075E+06	0.347756E+03	0.826998E+02
0.131000E+05	0.990434E+06	0.350693E+03	0.827328E+02
0.136000E+05	0.103181E+07	0.353530E+03	0.827622E+02
0.141000E+05	0.107320E+07	0.356274E+03	0.827886E+02
0.146000E+05	0.111460E+07	0.358930E+03	0.828124E+02
0.151000E+05	0.115601E+07	0.361504E+03	0.828338E+02
0.156000E+05	0.119743E+07	0.364002E+03	0.828533E+02
0.161000E+05	0.123886E+07	0.366426E+03	0.828709E+02
0.166000E+05	0.128030E+07	0.368782E+03	0.828870E+02
0.171000E+05	0.132175E+07	0.371074E+03	0.829017E+02
0.176000E+05	0.136320E+07	0.373304E+03	0.829152E+02
0.181000E+05	0.140466E+07	0.375475E+03	0.829276E+02
0.186000E+05	0.144613E+07	0.377592E+03	0.829390E+02
0.191000E+05	0.148760E+07	0.379656E+03	0.829496E+02
0.196000E+05	0.152908E+07	0.381671E+03	0.829593E+02

Thermodynamic Properties of C2-  
Table IV

T ( K )	H-HR (J/mol)	-(G-HR) / T (J/mol)	CP (J/mol)
0.100000E+03	0.190333E+04	0.145637E+03	0.291055E+02
0.600000E+03	0.167883E+05	0.189525E+03	0.315195E+02
0.110000E+04	0.339465E+05	0.207272E+03	0.371805E+02
0.160000E+04	0.537381E+05	0.219326E+03	0.415357E+02
0.210000E+04	0.750204E+05	0.228753E+03	0.432107E+02
0.260000E+04	0.966832E+05	0.236544E+03	0.432796E+02
0.310000E+04	0.118212E+06	0.243172E+03	0.428002E+02
0.360000E+04	0.139476E+06	0.248922E+03	0.422671E+02
0.410000E+04	0.160498E+06	0.253988E+03	0.418408E+02
0.460000E+04	0.181339E+06	0.258509E+03	0.415455E+02
0.510000E+04	0.202062E+06	0.262587E+03	0.413606E+02
0.560000E+04	0.222713E+06	0.266300E+03	0.412573E+02
0.610000E+04	0.243328E+06	0.267706E+03	0.412105E+02
0.660000E+04	0.263930E+06	0.272853E+03	0.412006E+02
0.710000E+04	0.284533E+06	0.275776E+03	0.412142E+02
0.760000E+04	0.305146E+06	0.278506E+03	0.412421E+02
0.810000E+04	0.325776E+06	0.281066E+03	0.412782E+02
0.860000E+04	0.346425E+06	0.283477E+03	0.413189E+02
0.910000E+04	0.367096E+06	0.285755E+03	0.413620E+02
0.960000E+04	0.387788E+06	0.287915E+03	0.414061E+02
0.101000E+05	0.408502E+06	0.289967E+03	0.414507E+02
0.106000E+05	0.429238E+06	0.291922E+03	0.414955E+02
0.111000E+05	0.449997E+06	0.293790E+03	0.415404E+02
0.116000E+05	0.470779E+06	0.295577E+03	0.415857E+02
0.121000E+05	0.491583E+06	0.297291E+03	0.416314E+02
0.126000E+05	0.512410E+06	0.298936E+03	0.416777E+02
0.131000E+05	0.533261E+06	0.300520E+03	0.417248E+02
0.136000E+05	0.554135E+06	0.302045E+03	0.417729E+02
0.141000E+05	0.575034E+06	0.303517E+03	0.418221E+02
0.146000E+05	0.595957E+06	0.304939E+03	0.418725E+02
0.151000E+05	0.616907E+06	0.306314E+03	0.419242E+02
0.156000E+05	0.637882E+06	0.307645E+03	0.419773E+02
0.161000E+05	0.658884E+06	0.308936E+03	0.420317E+02
0.166000E+05	0.679914E+06	0.310188E+03	0.420877E+02
0.171000E+05	0.700972E+06	0.311404E+03	0.421450E+02
0.176000E+05	0.722059E+06	0.312586E+03	0.422039E+02
0.181000E+05	0.743176E+06	0.313736E+03	0.422641E+02
0.186000E+05	0.764324E+06	0.314855E+03	0.423258E+02
0.191000E+05	0.785502E+06	0.315945E+03	0.423889E+02
0.196000E+05	0.806713E+06	0.317009E+03	0.424533E+02

Thermodynamic Properties of CO+  
Table V

T ( K )	H-HR (J/mol)	-(G-HR) /T (J/mol)	CP (J/mol)
0.100000E+03	-0.578048E+04	0.229123E+03	0.291041E+02
0.600000E+03	0.891734E+04	0.208895E+03	0.303617E+02
0.110000E+04	0.249584E+05	0.220413E+03	0.336090E+02
0.160000E+04	0.422639E+05	0.229635E+03	0.354052E+02
0.210000E+04	0.602218E+05	0.237134E+03	0.363347E+02
0.260000E+04	0.785381E+05	0.243426E+03	0.368935E+02
0.310000E+04	0.970945E+05	0.248839E+03	0.373240E+02
0.360000E+04	0.115864E+06	0.253587E+03	0.377620E+02
0.410000E+04	0.134870E+06	0.257820E+03	0.382784E+02
0.460000E+04	0.154159E+06	0.261640E+03	0.388935E+02
0.510000E+04	0.173778E+06	0.265127E+03	0.395935E+02
0.560000E+04	0.193761E+06	0.268338E+03	0.403463E+02
0.610000E+04	0.214126E+06	0.271319E+03	0.411145E+02
0.660000E+04	0.234872E+06	0.274103E+03	0.418633E+02
0.710000E+04	0.255982E+06	0.276719E+03	0.425650E+02
0.760000E+04	0.277426E+06	0.279188E+03	0.432000E+02
0.810000E+04	0.299169E+06	0.281527E+03	0.437569E+02
0.860000E+04	0.321169E+06	0.283752E+03	0.442310E+02
0.910000E+04	0.343386E+06	0.285873E+03	0.446228E+02
0.960000E+04	0.365779E+06	0.287901E+03	0.449364E+02
0.101000E+05	0.388310E+06	0.289845E+03	0.451785E+02
0.106000E+05	0.410947E+06	0.291710E+03	0.453569E+02
0.111000E+05	0.433658E+06	0.293504E+03	0.454799E+02
0.116000E+05	0.456419E+06	0.295232E+03	0.455558E+02
0.121000E+05	0.479207E+06	0.296897E+03	0.455925E+02
0.126000E+05	0.502006E+06	0.298506E+03	0.455971E+02
0.131000E+05	0.524800E+06	0.300061E+03	0.455759E+02
0.136000E+05	0.547578E+06	0.301565E+03	0.455347E+02
0.141000E+05	0.570332E+06	0.303022E+03	0.454781E+02
0.146000E+05	0.593055E+06	0.304435E+03	0.454103E+02
0.151000E+05	0.615741E+06	0.305805E+03	0.453347E+02
0.156000E+05	0.638389E+06	0.307136E+03	0.452542E+02
0.161000E+05	0.660995E+06	0.308429E+03	0.451709E+02
0.166000E+05	0.683559E+06	0.309687E+03	0.450869E+02
0.171000E+05	0.706082E+06	0.310910E+03	0.450036E+02
0.176000E+05	0.728563E+06	0.312102E+03	0.449222E+02
0.181000E+05	0.751005E+06	0.313263E+03	0.448400E+02
0.186000E+05	0.773407E+06	0.314395E+03	0.447685E+02
0.191000E+05	0.795774E+06	0.315499E+03	0.446975E+02
0.196000E+05	0.818106E+06	0.316576E+03	0.446310E+02

Thermodynamic properties of H3+  
Table VI

T ( K )	H-HR (J/mol)	-(G-HR) /T (J/mol)	CP (J/mol)
0.100000E+03	-0.658706E+04	0.195679E+03	0.332574E+02
0.600000E+03	0.101333E+05	0.172686E+03	0.342431E+02
0.110000E+04	0.286065E+05	0.185782E+03	0.400884E+02
0.160000E+04	0.501565E+05	0.196522E+03	0.458179E+02
0.210000E+04	0.741272E+05	0.205584E+03	0.498143E+02
0.260000E+04	0.997774E+05	0.213452E+03	0.526414E+02
0.310000E+04	0.126656E+06	0.220421E+03	0.547812E+02
0.360000E+04	0.154482E+06	0.226685E+03	0.564584E+02
0.410000E+04	0.183054E+06	0.232379E+03	0.577733E+02
0.460000E+04	0.212205E+06	0.237603E+03	0.587849E+02
0.510000E+04	0.241796E+06	0.242429E+03	0.595417E+02
0.560000E+04	0.271712E+06	0.246916E+03	0.600889E+02
0.610000E+04	0.301857E+06	0.251107E+03	0.604681E+02
0.660000E+04	0.332158E+06	0.255039E+03	0.607158E+02
0.710000E+04	0.362556E+06	0.258741E+03	0.608629E+02
0.760000E+04	0.393008E+06	0.262238E+03	0.609346E+02
0.810000E+04	0.423482E+06	0.265552E+03	0.609508E+02
0.860000E+04	0.453952E+06	0.268698E+03	0.609269E+02
0.910000E+04	0.484404E+06	0.271694E+03	0.608749E+02
0.960000E+04	0.514824E+06	0.274552E+03	0.608037E+02
0.101000E+05	0.545206E+06	0.277284E+03	0.607200E+02
0.106000E+05	0.575543E+06	0.279900E+03	0.606287E+02
0.111000E+05	0.605834E+06	0.282409E+03	0.605335E+02
0.116000E+05	0.636076E+06	0.284820E+03	0.604369E+02
0.121000E+05	0.666271E+06	0.287139E+03	0.603409E+02
0.126000E+05	0.696417E+06	0.289372E+03	0.602467E+02
0.131000E+05	0.726518E+06	0.291527E+03	0.601552E+02
0.136000E+05	0.756573E+06	0.293608E+03	0.600671E+02
0.141000E+05	0.786585E+06	0.295619E+03	0.599825E+02
0.146000E+05	0.816556E+06	0.297565E+03	0.599018E+02
0.151000E+05	0.846488E+06	0.299451E+03	0.598250E+02
0.156000E+05	0.876382E+06	0.301279E+03	0.597520E+02
0.161000E+05	0.906241E+06	0.303053E+03	0.596829E+02
0.166000E+05	0.936066E+06	0.304776E+03	0.596174E+02
0.171000E+05	0.965859E+06	0.306451E+03	0.595554E+02
0.176000E+05	0.995622E+06	0.308080E+03	0.594969E+02
0.181000E+05	0.102536E+07	0.309666E+03	0.594415E+02
0.186000E+05	0.105506E+07	0.311211E+03	0.593892E+02
0.191000E+05	0.108475E+07	0.312716E+03	0.593397E+02
0.196000E+05	0.111440E+07	0.314185E+03	0.592929E+02



Thermodynamic Properties of H<sub>2</sub>O+  
Table VII

T ( K )	H-HR (J/mol)	-(G-HR) / T (J/mol)	CP (J/mol)
0.100000E+03	-0.668326E+04	0.149109E+03	0.332574E+02
0.600000E+03	0.104914E+05	0.125531E+03	0.368439E+02
0.110000E+04	0.306992E+05	0.139387E+03	0.439456E+02
0.160000E+04	0.541624E+05	0.150961E+03	0.495773E+02
0.210000E+04	0.799952E+05	0.160745E+03	0.535230E+02
0.260000E+04	0.107491E+06	0.169230E+03	0.562955E+02
0.310000E+04	0.136150E+06	0.176730E+03	0.582227E+02
0.360000E+04	0.165614E+06	0.183454E+03	0.595497E+02
0.410000E+04	0.195631E+06	0.189550E+03	0.604629E+02
0.460000E+04	0.226030E+06	0.195123E+03	0.610927E+02
0.510000E+04	0.256691E+06	0.200255E+03	0.615251E+02
0.560000E+04	0.287531E+06	0.205011E+03	0.618156E+02
0.610000E+04	0.318489E+06	0.209439E+03	0.620010E+02
0.660000E+04	0.349519E+06	0.213582E+03	0.621070E+02
0.710000E+04	0.380586E+06	0.217473E+03	0.621522E+02
0.760000E+04	0.411664E+06	0.221141E+03	0.621508E+02
0.810000E+04	0.442731E+06	0.224608E+03	0.621139E+02
0.860000E+04	0.473773E+06	0.227895E+03	0.620502E+02
0.910000E+04	0.504778E+06	0.231019E+03	0.619668E+02
0.960000E+04	0.535738E+06	0.233995E+03	0.618693E+02
0.101000E+05	0.566646E+06	0.236836E+03	0.617621E+02
0.106000E+05	0.597499E+06	0.239553E+03	0.616488E+02
0.111000E+05	0.628294E+06	0.242157E+03	0.615320E+02
0.116000E+05	0.659031E+06	0.244655E+03	0.614138E+02
0.121000E+05	0.689708E+06	0.247057E+03	0.612960E+02
0.126000E+05	0.720327E+06	0.249368E+03	0.611796E+02
0.131000E+05	0.750888E+06	0.251596E+03	0.610656E+02
0.136000E+05	0.781393E+06	0.253746E+03	0.609547E+02
0.141000E+05	0.811843E+06	0.255822E+03	0.608473E+02
0.146000E+05	0.842241E+06	0.257831E+03	0.607437E+02
0.151000E+05	0.872588E+06	0.259775E+03	0.606440E+02
0.156000E+05	0.902886E+06	0.261659E+03	0.605483E+02
0.161000E+05	0.933137E+06	0.263486E+03	0.604567E+02
0.166000E+05	0.963343E+06	0.265260E+03	0.603692E+02
0.171000E+05	0.993506E+06	0.266983E+03	0.602855E+02
0.176000E+05	0.102363E+07	0.268658E+03	0.602057E+02
0.181000E+05	0.105371E+07	0.270288E+03	0.601296E+02
0.186000E+05	0.108376E+07	0.271876E+03	0.600570E+02
0.191000E+05	0.111377E+07	0.273422E+03	0.599879E+02
0.196000E+05	0.114375E+07	0.274929E+03	0.599220E+02

# **NON-INTRUSIVE TESTING OF COMPOSITE AIRCRAFT ENGINE COMPONENTS: I**

**Laurence J. Jacobs, Assistant Professor**

**Engineering Science and Mechanics Program**

**Georgia Institute of Technology**

## **ABSTRACT**

Advanced, high temperature composite engine components will add new complications to established testing procedures and increase the need for innovative non-intrusive technologies. Lightweight composite materials are inhomogeneous and anisotropic in nature; these conditions complicate any potential structural model and question the validity of surface, dynamic stress measurements. Laser ultrasonics has the capability to provide useful engineering information about the structural integrity of an in-service engine component without interfering with the process being monitored. Since elastic waves propagate through the component thickness, laser ultrasonics examines the condition of the entire specimen and not just a limited number of surface points. Ultrasonics is a powerful tool for the characterization of structural materials, but to be effective, the propagation characteristics of the elastic waves themselves must be understood. This report concludes with a review of the basic aspects of wave propagation in an anisotropic media and discusses the potential for material characterization using laser ultrasonics.

## INTRODUCTION

The use of composite materials for propulsion systems will dramatically increase the need for new technologies for engine testing and certification. There are a variety of potential problems if methodologies currently in use for "traditional" engine materials are applied to these new, high performance materials. The purpose of this report is to examine the potential of innovative, non-intrusive techniques to monitor the structural integrity of propulsion components. Of primary concern is the ability of the methodology to measure useful engineering values which can be related to the component's integrity. Part I of this report introduces the proposed system, laser ultrasonics, and discusses in depth, its application to a potential advanced composite engine material, Avimid-N. The second part of this report (authored by graduate student James Abbey) summarizes other potential nondestructive evaluation techniques, examines, in depth, the laser generation of ultrasound and reviews the candidate composite material types.

## PROBLEM STATEMENT AND PROPOSED SOLUTION

Lightweight, high temperature composite engine materials have the potential to provide significant economic and performance benefits. There are, however, many problems associated with the use of these new materials; chief among them is that the structural behavior of a composite can be very different from that of a metal component. A composite's inherent anisotropy governs both its static and dynamic structural behavior. This anisotropy, which is tailored by the designer to obtain maximum structural performance, must be properly accounted for in any

evaluation procedure.

The current engine testing technology, which is very well established and has been field proven, uses dynamic strain gages to directly obtain measurements of certain critical parts. There are significant potential problems concerning the application of this technique for composite components. The primary difficulty is that strain gages make surface measurements at a limited number of points; these values are used, with the appropriate structural analysis, to infer the state of stress throughout the component. The limited life of the gages, their associated wire paths and the expense involved in their installation are additional disadvantages for their use with these new composite materials. The effectiveness of a surface measurement of a composite material is debatable because of the material's anisotropic behavior [1]. Additionally, detailed reliable analysis packages are not currently available for these advanced materials, so the usefulness of a limited number of single point strain measurements is doubtful for composite engine components.

Laser ultrasonics [2] is a non-contact, non-intrusive, nondestructive evaluation technique that uses lasers to generate and detect ultrasonic waves. The physical principles underlying the generation and reception of laser ultrasound in metallic specimens is fairly well understood. Briefly, a certain amount of the incident electromagnetic energy from the laser is absorbed at the surface of the specimen and converted into thermal energy. The nanosecond time scale of a Q-switched laser pulse causes the thermal wavefield to extend only a few millimeters into the specimen. This nearly instantaneous thermal expansion generates elastic waves which propagate into the specimen. The specimen's response to these elastic waves

is measured with a laser interferometer. This optical device permits the high fidelity localized detection of ultrasonic waves without disturbing the process being monitored. A schematic of the laser ultrasonic process is given in Figure 1. Part II of this report provides a detailed description of the laser generation of ultrasound and its application to high temperature composite materials.

The main advantages of laser ultrasonics for the non-intrusive testing of composite engine components include:

- (1) It is a totally non-contact technique, so it does not interfere with the process being monitored. In addition, the adverse temperature and vibration environment does not cause excessive difficulties. There is also no need for complicated preparation or time consuming pre-test installations.
- (2) The method provides information about the condition of the material throughout the thickness of the component, not just at a single point on its surface. This information includes wave velocities, attenuation and dispersion which are used to calculate meaningful, engineering data on the structural integrity of the component such as temperature variation and elongation.
- (3) The generation and detection of ultrasound is accomplished using fiber optic cables. These leads allow for easy access into the test cell and enables the system to interrogate difficult to reach engine components during testing.

At the present time, laser ultrasonics has proven to be an effective tool for the laboratory monitoring of metallic components. An enormous effort is required in order for it to make the transition from a laboratory technique to the in-service monitoring of rotating, composite engine components. This

report summarizes efforts to establish the analytical fundamentals for the application of laser ultrasonics to a high temperature, organic matrix composite material, Avimid-N. The next step, which is the subject of the author's proposed follow-up mini-grant, involves benchmark laboratory experiments to establish the viability of the laser generation and quantitative detection of ultrasound in Avimid-N. Samples of this material have already been procured from the manufacturer (see Appendix) for this proposed, experimental study.

## LASER DETECTION OF ULTRASOUND

It is possible to combine the laser generation of ultrasound with piezoelectric transducers, electromagnetic acoustic transducers (EMAT) or capacitive transducers, but a truly non-intrusive system is only established when a remote, optical system [3,4] is used to detect the ultrasonic waves; this optical system is the only viable option for the non-intrusive testing of aircraft engine components. An example of a heterodyne interferometer is shown in Figure 2. This optical system is operated by splitting a single frequency laser light into two components using an acousto-optic modulator. These two components, which are separated in frequency by 40MHz, are sent along two arms of an interferometer, one of which contains the sample being monitored. The face of the specimen serves as one mirror in the interferometer. The beams are recombined on the surface of a photodetector producing a beat frequency of 40MHz. Phase shifts in the light reflected from the sample's surface result in proportional phase shifts in the beat signal. As a result, the 40MHz signal

acts as a carrier that can be demodulated to determine the time dependent velocity occurring at the sample's surface. The signal at the photodetector can be demodulated in real time using an FM discriminator. The demodulated output signal is proportional to the normal surface velocity of the specimen and can be integrated to determine its time dependent displacement.

## ELASTIC WAVE PROPAGATION CONSIDERATIONS

Ultrasonics is a very powerful tool for the characterization of structural materials. In order to use ultrasonic waves to effectively probe a sample, the propagation characteristics of the ultrasonic waves themselves must be fully understood. It is necessary to review the basic aspects of wave propagation in anisotropic materials. Specifically, it is critical to determine what can be physically measured and what can be learned from the ultrasonic evaluation of a composite specimen. Here the anisotropic nature of the composite material plays the central role. An isotropic material, such as most metals, has only two independent elastic constants and does not possess any directional preferences; elastic wave propagation in isotropic materials is fairly well established [5]. In contrast, the most general anisotropic material has twenty-one independent stiffness (elastic) constants; now a strong directional preference exists. In recent years, there has been a great deal of interest in wave propagation in anisotropic material in seismology [6,7], crystal acoustics [8] and electromagnetic fields [9]. In general, three different linear elastic waves may propagate along any given direction in an anisotropic material and each of these waves can travel with distinct phase velocities. An additional complication

is that the energy-velocity vectors and the phase-velocity vectors for plane wave propagation are the same for an isotropic material, while in general, they are different for an anisotropic material. This means that in an anisotropic material, the acoustic energy does not propagate along the wave front direction; this difference will greatly affect any ultrasonic measurements made.

In general, a layered composite component can be modeled as an orthotropic material; here nine independent elastic constants are required to specify the stress-strain relationship. An additional level of symmetry, transverse isotropy, exists for a unidirectional (all of the reinforcing fibers are aligned in a single direction) composite specimen and reduces the number of independent elastic constants to five. Lightweight, fiber reinforced composites are not homogeneous materials. However, most approaches to the mechanics of composite materials begin with the simplifying assumption that the composite material may be treated as homogeneous. For the ultrasonic evaluation of composites, if the layer thickness and fibers are small compared to the wavelengths of the ultrasonic waves, the inherent inhomogeneity of the fiber reinforced composite can be ignored and the material treated as homogeneous.

The time harmonic equations of motion for a general anisotropic media reduce to

$$\det|\Gamma_{ij} - \rho c^2 \delta_{ij}| = \Omega = f(\omega, k_x, k_y, k_z) = 0 \quad (1)$$

Here the constants  $\Gamma_{ij}$  are known as the Christoffel stiffnesses, and can be written in terms of the material's twenty-one independent elastic stiffness components.



Equation (1) describes three velocity fronts in space for any given direction. The physical interpretation of this means that for any direction of wave propagation, there are three distinct phase velocities and their three corresponding displacement values are orthogonal. As opposed to the isotropic case, these displacements are neither truly longitudinal nor truly transverse. These phase velocities are best presented in terms of their inverse values, or slownesses. For example, examine the elastic wave propagation in a transversely isotropic material (the model for a unidirectional composite). Equation (1) reduces to calculating the determinant of a three by three system of equations. For the simplest direction of propagation, perpendicular to the axis of symmetry (the fiber direction), the wave speeds are independent of the direction of propagation in that plane only. This is shown in a plot of their slowness curves (inverse of the phase velocities), given in Figure 3.

Instead of traveling with the phase velocity, the energy of the ultrasonic waves in an anisotropic media will travel with the energy velocity, which is identical to the group velocity (in a lossless media). This difference must be accounted for in the design of an effective ultrasonic evaluation experiment in an anisotropic medium. The group velocity vector, and thus the energy velocity vector, is calculated from the dispersion relation, equation (1), by

$$\mathbf{V}_e = \mathbf{V}_g = -\frac{\nabla_{\mathbf{k}} \Omega}{\partial \Omega / \partial \omega} \quad (2)$$

The group velocity components are the values that are physically measured and serve as the basis for the calculation of the phase velocity (or slowness) components. The phase velocities are more useful, since the solution to all problems of wave

propagation are tied to their values.

In addition to the bulk wave velocity considerations just discussed, the possibility of using guided waves [10-14] for the ultrasonic inspection of composite materials must be considered. These plate waves may provide useful information for the analysis of thin composite layers. Here, structural vibration considerations dominate and values of resonance frequencies and dispersion curves are indirectly used for material characterization. The use of these plate waves depends more on the component geometry (plate thickness and boundary conditions) and occur on a slower time scale than the bulk waves. Dispersion curves (frequency versus phase velocity) are available [12,13] and can be incorporated into any testing scheme.

#### CALCULATION OF ELASTIC CONSTANTS

The required starting point for the application of laser ultrasonics to the characterization of high temperature, composite aircraft engine materials is the experimental validation of the technique for the material being considered, Avimid-N. The answer to the question of whether laser generation and detection is possible with this material is best determined in a laboratory environment. In the laboratory, the first step in the ultrasonic inspection of an anisotropic material is to accurately acquire its material properties, specifically, the matrix of its elastic constants which describe the homogeneous medium of a perfectly manufactured material [15-17]. The case of a unidirectionally reinforced composite, modeled as a transversely isotropic material, with  $z$  as the axis of symmetry (as shown in Figure 4), comprises determining the five independent elastic constants

$$(C_{ij}) = \begin{pmatrix} C_{11} & C_{12} & C_{23} & 0 & 0 & 0 \\ C_{12} & C_{22} & C_{23} & 0 & 0 & 0 \\ C_{23} & C_{23} & C_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{55} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & C_{66} \end{pmatrix} \quad (3)$$

where  $C_{12} = C_{11} - 2C_{66}$ .

This is the proposed symmetry model for unidirectional, carbon reinforced, Avimid-N. It is important to note that these elastic constants directly depend upon the phase velocities, as given in equation (1).

Besides being a group of material constants, these stiffness components (equation (3)) contain information about the reference state of the material in a particular neighborhood and can be used to determine changes from this reference state. Once the reference state of the material is completely characterized, the ultrasonic waves can be used to determine defects such as voids, delaminations and porosity as well as potentially detect some more subtle changes like temperature variation, elongation and state of stress. These differences, which will appear as changes in the measured ultrasonic waveform's shape, phase velocity, attenuation and dispersion will be discussed in detail in the next section. A more empirical approach to material characterization using ultrasonic waves, the stress wave factor, is discussed in part II of these report.

The anisotropy of the material causes enormous complications in any ultrasonic investigation. The fundamental aspects of wave propagation are effected, including energy velocity, beam splitting, beam skewing, unsymmetrical field profiles, unusual side lobes and beam divergence. These can all cause errors in accurate material characterization. This heightens the importance of understanding anisotropic wave propagation, as discussed in the previous section.

The experimentally obtained phase velocities are verified using the previously determined numerical results, which account for the difference between energy and phase velocity directions. This involves the solution of a system of nonlinear algebraic equations for the unknown phase velocities. These results are used to fill in the elastic constant matrix of equation (3). A tentative procedure for the experimental determination of the elastic constants of a transversely isotropic (unidirectional) composite material is:

- (1) Measure the longitudinal phase velocity along the fiber direction to evaluate constant  $C_{33}$ .
- (2) Measure the longitudinal phase velocity perpendicular to the fiber direction to evaluate constants  $C_{11}$  and  $C_{22}$ .
- (3) Measure the vertically polarized shear velocity perpendicular to the fibers to evaluate constants  $C_{44}$  and  $C_{55}$ .
- (4) Measure the horizontally polarized shear velocity along the fiber direction to evaluate  $C_{66}$  (calculate  $C_{12} = C_{11} - 2C_{66}$ ).
- (5) Measure the quasi-longitudinal phase velocities at various angles to the fiber direction to evaluate  $C_{13}$  and  $C_{23}$ .

So, by taking advantage of the symmetry of the transversely isotropic composite and varying the incident angle, redundancies can be utilized to check the validity of the measurements.

## MATERIAL CHARACTERIZATION USING ULTRASONICS

The overall stiffness matrix of equation (3) provides a frame of reference to begin quantitative material characterization. This is the most critical application for aircraft engine testing applications. Since laser ultrasonics provides a very broad band signal, it is particularly well suited for detecting localized changes in elastic properties. Traditional ultrasonics has been used to detect changes in temperature [18], microstructure [19] and to measure residual stresses [20]. The technique has proven to be quite effective in characterizing and monitoring a material's microstructure, even during fabrication processes. Because the ultrasonic waves propagate through the thickness of the specimen, laser ultrasonics provides information about the condition of the entire component.

The relationship between applied stress and the resultant strain in a solid material is usually considered to be linear. But even within the elastic range, small departures from linearity exist; this means that the elastic constants vary slightly with stress. The variation of the constants' values can be very small, on the order of fractions of a percent. Acoustoelasticity [20] uses elastic (ultrasonic) waves to quantify these small changes in stress. As shown in the previous section, the phase velocity of elastic waves in a medium are dependent on the elastic constants, so it is possible to determine the state of stress using accurate measurements of changes

in wave velocities. Acoustoelasticity is based on a continuum theory of small disturbances (the ultrasonic waves) superimposed on an elastically deformed body (the engine component). The advantage of laser ultrasonics is the accuracy of its measurements and its repeatability, which allows for the use of time averaging schemes. This is an area that demands further research, after the experimental measurement of Avimid-N's "reference" elastic constants.

The temperature variation through the thickness is not based on a direct measurement; instead temperature dependent physical properties are used. Again, accurate measurement of the elastic constants is necessary. The laser ultrasonic method is capable of making the required time-of-flight (phase velocity) measurements. Extreme caution is required since, like the acoustoelastic effect, the expected changes are very small and are dependent on other factors such as microinhomogeneity and texture. For these reasons, the use of the ultrasonic technique to measure initial stresses and temperature variation is still being developed. This does not preclude the application of laser ultrasonics for the non-intrusive evaluation of composite engine components; the technique is unmatched in its potential for this purpose.

## CONCLUSION

Laser ultrasonics is a powerful tool for the characterization of structural materials. The method can provide relevant engineering data about the structural integrity of the entire component without interfering with the process being monitored. Its potential to investigate in-service, high temperature composite

materials is unmatched. To be effective, the propagation characteristics of elastic waves in an anisotropic media must be examined. The waveform, phase velocity, attenuation and dispersion of these ultrasonic waves contain useful information about the structural integrity of an engineering component. The required first step for the application of laser ultrasonics to the characterization of high temperature, organic composite materials is the laboratory determination of the candidate material's, Avimid-N, elastic constants (as given in equation (3)). This will not only validate the laser ultrasonic method for Avimid-N, but also provide information about the reference state of this material in a particular neighborhood. Once this reference state is completely characterized, elastic waves can be used to detect subtle changes from this reference state such as elongation and temperature variation. It is these values that provide useful engineering information to test and analysis engineers.

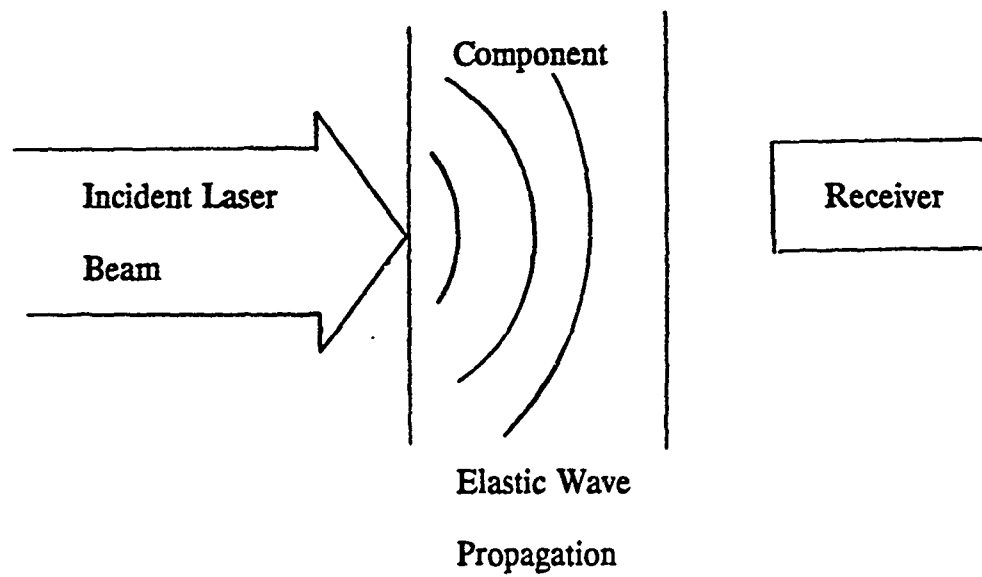


Figure 1. Schematic of Laser Ultrasonics

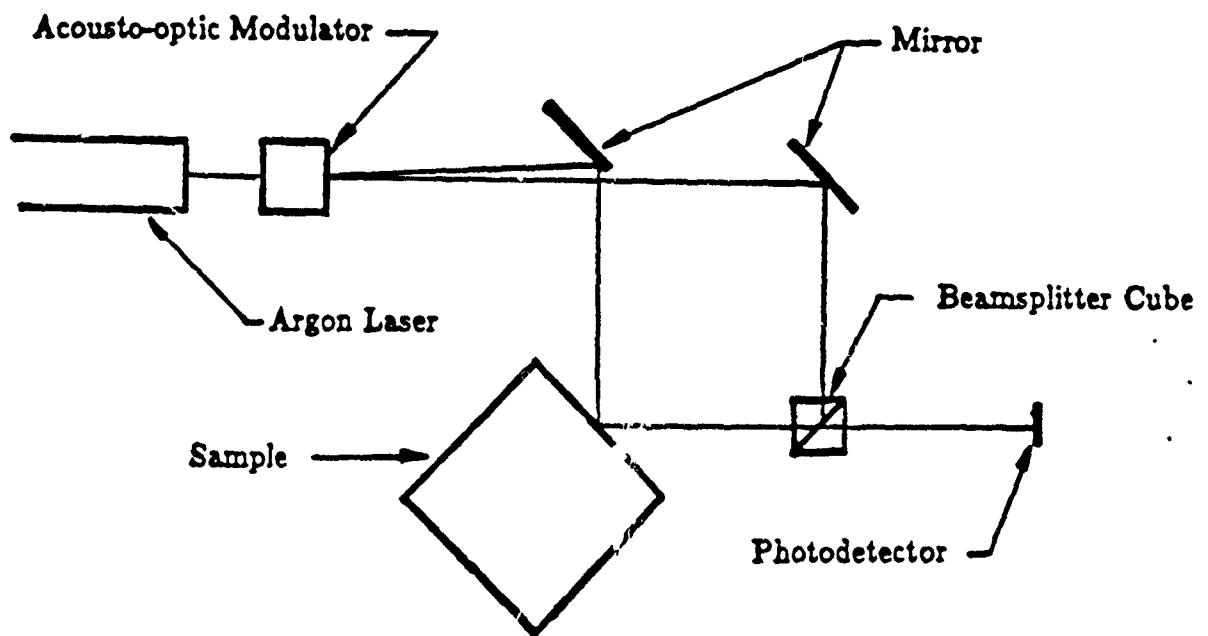


Figure 2. Interferometer



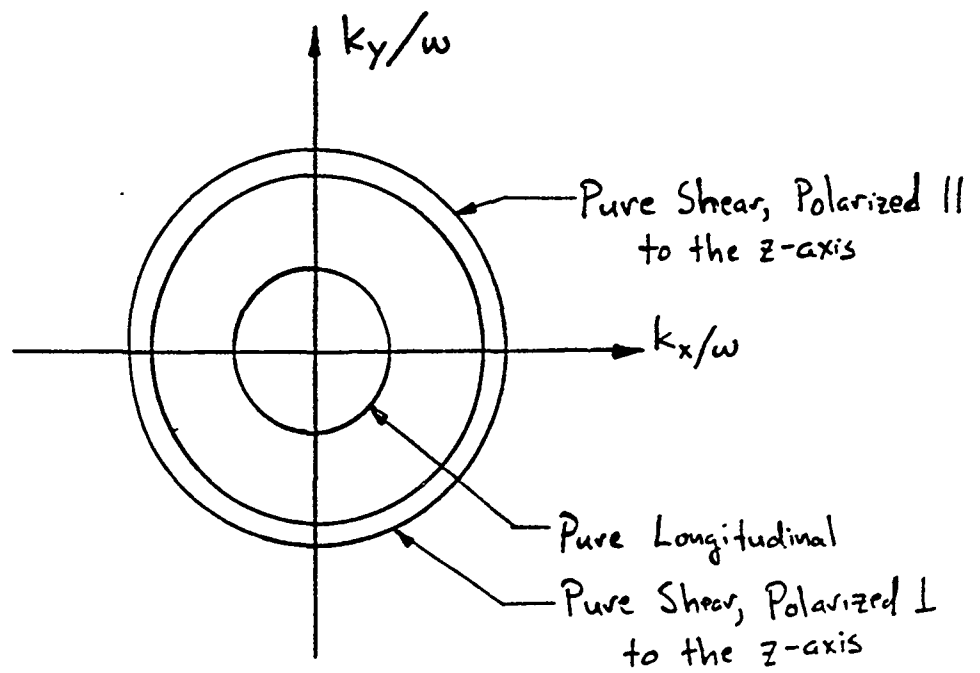


Figure 3. Typical Slowness Curve

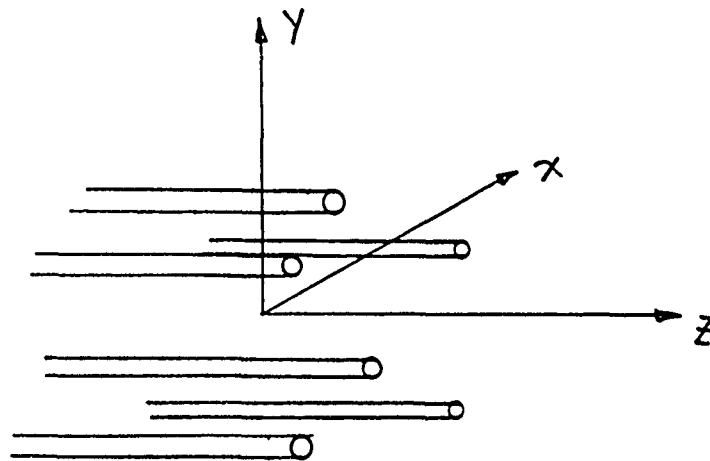


Figure 4. Unidirectional Composite

## REFERENCES

1. M.E. Tuttle and H.F. Brinson, "Resistance foil strain gage technology as applied to composite materials," Experimental Mechanics, Vol. 24, pp. 54-65, 1984.
2. C.B. Scruby, "Some applications of laser ultrasound," Ultrasonics, Vol. 27, pp. 195-209, 1989.
3. J.W. Wagner, "Optical detection of ultrasound," Physical Acoustics, W.P. Mason and R.N. Thurston, eds., Vol. 19, 1990. 4. J.P. Monchalin, J.D. Aussel, R. Heon, C.K. Jen, A. Boudreault and R. Bernier, "Measurement of in-plane and out-of-plane ultrasonic displacements by optical heterodyne interferometry," Journal of Nondestructive Evaluation, Vol. 8, pp. 121-133, 1989.
5. J.D. Achenbach, Wave Propagation in Elastic Solids, North-Holland, 1973.
6. K. Aki and P.G. Richards, Quantitative Seismology Theory and Methods, Vols. 1 and 2, W.H. Freeman, 1980.
7. J.H.M.T. Van der Hijden, Propagation of Transient Elastic Waves in Stratified Anisotropic Media, North-Holland, 1987. 8. M.J.P. Musgrave, Crystal Acoustics, Holden-Day, 1970.
9. B.A. Auld, Acoustic Fields and Waves in Solids, Vols. 1 and 2, 2nd edition, R.E. Krieger, 1990.
10. I.A. Viktorov, Rayleigh and Lamb Waves, Plenum Press, 1967.
11. J.B. Spicer, A.D.W. McKie and J.W. Wagner, "Quantitative theory for laser ultrasonic waves in a thin plate," Applied Physics Letters, Vol. 29, pp. 1882-1884, 1990.

12. Y.Li and R.B. Thompson, "Influence of anisotropy on the dispersion characteristics of guided ultrasonic plate modes," Journal of the Acoustical Society of America, Vol. 87, pp. 1911-1931, 1990.
13. A.H. Nayfeh and D.E. Chimenti, "Free wave propagation in plates of general anisotropic media," Journal of Applied Mechanics, Vol. 56, pp. 881-886, 1989.
14. R.C. Stiffler and E.G. Henneke, "Low frequency plate modes," NASA Contractor Report 1976, Ultrasonic Stress Wave Characterization of Composite Materials, May, 1986.
15. R.A. Kline and Z.T. Chen, "Ultrasonic technique for global anisotropic measurement in composite materials," Materials Evaluation, Vol. 46, pp. 986-992, 1988.
16. J.L. Rose, A. Pilarski, K. Balasubramaniam, A. Tverdokhlebov and J. Ditri, "Ultrasonic wave considerations for the deployment in an NDE feature matrix for anisotropic media," Journal of Engineering Materials and Technology, Vol. 111, pp. 225-262, 1989.
17. T.T. Wu and Z.H. Ho, "Anisotropic wave propagation and its application to NDE of composite materials," Experimental Mechanics, Vol. 30, pp. 313-318, 1990.
18. H.N.G. Wadley, S.J. Norton, F. Mauer and B. Droney, "Ultrasonic measurement of internal temperature distribution," Philosophical Transactions: Royal Society of London, Vol. A320, pp. 341-361, 1986.
19. C.B. Scruby, R.L. Smith and B.C. Moss, "Microstructural monitoring by laser ultrasonic attenuation and forward scattering," NDT International, Vol. 19, pp.307-313, 1986.

20. Y.H. Pao, W. Sachse and H. Fukuioka, "Aconstoelasticity and ultrasonic measurements of residual stress," Physical Acoustics, W.P. Mason and R.N. Thurston, eds., Vol. 17, pp. 61-143, 1984.

## **APPENDIX**

### **Key Phone Conversations**

#### **1. Dr. Dick Cornelia, Dupont (302) 999-2498**

**He is providing us with sample Avimid-N (high temperature, polyimide) specimens. This material should be used by Pratt and Whitney, reinforced with carbon fibers.**

#### **2. Dr. Tom Harding, Dupont**

**Discussed Dupont's NDE procedures for the Avimid-N. Currently using ultrasonic immersion techniques for defect detection.**

#### **3. Mr. Dick Froom, McClelland Air Force Base, California**

**(916) 643-4274 (commercial), 633-4274 (autovon)**

**He is developing a laser ultrasonic scanning system for the Air Force. Two systems are available, one that is capable of interrogating an entire aircraft and a second that looks at individual components.**

#### **4. Henry Jones, AEDC (454-7750)**

**Discussed accessibility issue of the fiber optic probes in an engine. Manufacturer will provide access holes (at locations of their choosing) through the engine case. Provided helpful information based on his experience with laser monitoring of blade tip deflections.**

#### **5. Carl Brasier, AEDC (454-4661)**

**Discussed current laser/optical capabilities available at Arnold. There exists, in house, sufficient hardware, including lasers, optical components and data acquisition systems for any future follow-on demonstration at AEDC.**

## IMPLEMENTATION OF MULTIGRID IN THE PARC CODE

Steven M. McKay

**Abstract.** The PARC code is a general purpose Navier Stokes solver developed at Arnold Engineering Development Center at Arnold Air Force Base, Tennessee. It is used continuously in design and analysis of flow characteristics of aircraft. However, these types of flows are very complicated and difficult to solve. Time dependent multigrid has been implemented in the PARC code which accelerates convergence to steady state. Convergence of the multigrid solver is on the order of convergence of the coarsest grid.

### 1. Introduction.

In the past decade computational speed and memory capacity has increased greatly. This has allowed for the solution of many large scale problems which previously were considered intractable. However, there are still a large class of problems for which solution is obtained only after great expense in terms of both computer time and man hours. Hence, there is a critical need to apply a solution technique which is efficient in terms of memory and provides a fast solution.

One such problem is that of the development of an optimal free-jet forebody design. Due to the nature of the problem, optimization of a free-jet forebody design simulator can only be accomplished by use of direct function optimization. In this case, the optimization is not cast as part of the fluid flow analysis. Instead measurements are made from a CFD solution of the fluid flow to determine the change in design. This type of optimization is naturally very expensive as it requires many calculations of the fluid flow to obtain a solution. Therefore an improvement in the speed of such solutions would greatly benefit

the above project.

Multigrid is a technique used to accelerate convergence of solutions of partial differential equations. Traditionally, multigrid has been used in the acceleration of the convergence of iterative matrix solvers. Recently good success has been achieved by applying multigrid like algorithms to the Navier Stokes equations and accelerating convergence to steady state. Multigrid can provide dramatic acceleration of the convergence process thus yielding a more efficient solution.

Another advantage of multigrid is the ability to use solvers already existing in code to find a solution. Multigrid can be thought of as an outer algorithm; the solver need not be specified. Though some compensation must be made for different solvers, this aspect of multigrid lends itself to robust code.

In this paper we describe the implementation of multigrid in the PARC code. We will show that there can be a tremendous savings in cost by use of the described algorithm. In section 2 we give a brief introduction to the Navier - Stokes equations. We describe the PARC code in section 3, and the multigrid algorithm in section 4. In section 5 we give a sample problem, and in section 6 we give our conclusions.

## 2. The Navier - Stokes Equations.

By applying certain assumptions to fluid flow through a body, the Navier - Stokes equations can be derived. (For greater detail, we refer the reader to [1]). For two dimensional flow, they can be written in vector form as

$$(1) \quad \frac{\partial Q}{\partial t} + \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2} = \frac{1}{Re} \left( \frac{\partial G_1}{\partial x_1} + \frac{\partial G_2}{\partial x_2} \right)$$

where

$$Q = \begin{pmatrix} \rho \\ \rho u_1 \\ \rho u_2 \\ E \end{pmatrix}, F_1 = \begin{pmatrix} \rho u_1 \\ \rho u_1^2 + P \\ \rho u_1 u_2 \\ (E + P) u_1 \end{pmatrix}, F_2 = \begin{pmatrix} \rho u_2 \\ \rho u_1 u_2 \\ \rho u_2^2 + P \\ (E + P) u_2 \end{pmatrix},$$

$$G_1 = \begin{pmatrix} 0 \\ \tau_{11} \\ \tau_{21} \\ u_1 \tau_{11} + u_2 \tau_{12} - q_1 \end{pmatrix}, G_2 = \begin{pmatrix} 0 \\ \tau_{12} \\ \tau_{22} \\ u_1 \tau_{21} + u_2 \tau_{22} - q_2 \end{pmatrix}$$

and where

$$\tau_{ij} = \mu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \lambda \left( \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} \right) \delta_{ij}$$

is the viscous stress tensor and

$$q_j = -\frac{K}{\beta_r Pr} \frac{\partial T}{\partial x_j}$$

is the heat flux vector.

Note that there are four equations and six variables of interest. In order to simplify the

above equations, certain assumptions are made.



First, Stokes hypothesis,

$$\lambda = -\frac{2}{3}\mu$$

is used in  $\tau_{ij}$ . Next, the PARC

code makes the assumption that the fluid is thermally and calorically perfect. Pressure and temperature are then calculated by the following equations:

$$P = (\gamma - 1) \left( E - \frac{1}{2} \rho (u_1^2 + u_2^2) \right)$$
$$T = \gamma (\gamma - 1) \left( E / \rho - \frac{1}{2} (u_1^2 + u_2^2) \right)$$

$\gamma$  is the ratio of specific heats. The above equations relate the pressure and temperature to the conservation variables and reduces (1) to a system of four equations with four unknowns. The difficulty in the Navier - Stokes equations is in the nonlinear relationships of the conservation variables. This leads to systems which are difficult to solve.

### 3. The PARC code.

The PARC code is a general Navier - Stokes solver. The current implementation has three main areas of interest, namely, implementation of boundary conditions, types of solvers, and grid blocking.

Boundary conditions are currently implemented in the PARC code to allow for internal or external flows, solid boundaries of different types, and symmetry boundaries. These have been implemented for subsonic and supersonic flow. All boundary conditions currently implemented are done via direct solution algorithms. That is, the conservation variables at the boundary are solved using only information from the previous solution. The

large number of boundary conditions available lead to a versatile code which can solve a large variety of problems.

There are two solvers currently implemented in the PARC code. These are an implementation of the Beam-Warming algorithm, and pseudo Runge-Kutta. The Beam-Warming algorithm implemented is Euler backward differencing. The system of equations can be written as

$$\Delta Q + \Delta t \left( \frac{\partial F_1^{n+1}}{\partial x_1} + \frac{\partial F_2^{n+1}}{\partial x_2} - \frac{1}{Re} \left( \frac{\partial G_1^{n+1}}{\partial x_1} + \frac{\partial G_2^{n+1}}{\partial x_2} \right) \right) = 0$$

where  $\Delta Q = Q^{n+1} - Q^n$  and  $\Delta t = t^{n+1} - t^n$ .  $F_j^{n+1}$  is approximated by a first order

taylor polynomial  $F_j^{n+1} \approx F_j^n + A_j^n \Delta Q$  where  $A_j = \frac{\partial F_j}{\partial Q_j}$ . The viscous flux vectors are

then lagged, i.e.,  $G^{n+1} \approx G^n$ . This leads to

$$(2) \quad \left( I + \Delta t \left( \frac{\partial}{\partial x_1} A_1 + \frac{\partial}{\partial x_2} A_2 \right) \right) \Delta Q = - \Delta t \left( \frac{\partial F_1^n}{\partial x_1} + \frac{\partial F_2^n}{\partial x_2} - \frac{1}{Re} \left( \frac{\partial G_1^n}{\partial x_1} + \frac{\partial G_2^n}{\partial x_2} \right) \right).$$

The derivatives in (2) are approximated by central difference operators. This leads to a complex system of equations. In order to alleviate the difficulty presented by such a matrix, the left hand side is approximately factored and an ADI algorithm is used to find a solution.

Because central differencing is used, artificial viscosity is required for stability. We refer the reader to [2] for a discussion of the artificial viscosity used. For a complete discussion on the Beam - Warming algorithm, see [1], pp. 489-494.

As can be seen from the above, Beam Warming is an implicit solver; the equations of the conservation variables involve values at neighboring points. The pseudo Runge-Kutta algorithm, however, is an explicit solver; the solution of the conservation variables depends only on the solution at previous time steps. This provides a less costly solution. However, the allowable time step is limited by the CFL condition to maintain stability.

Pseudo Runge-Kutta is in fact not related to the Runge-Kutta method for solution of ordinary differential equations. It is in fact a multistep time integration scheme. In the PARC code three, four or five stage schemes are implemented.

The five stage scheme is as follows:

$$\begin{aligned}
 Q(0) &= Q^n \\
 Q(1) &= Q(0) + \frac{1}{4} \Delta t R(Q(0)) \\
 Q(2) &= Q(1) + \frac{1}{6} \Delta t R(Q(1)) \\
 Q(3) &= Q(2) + \frac{3}{8} \Delta t R(Q(2)) \\
 Q(4) &= Q(3) + \frac{1}{2} \Delta t R(Q(3)) \\
 Q^{n+1} &= Q(4) + \Delta t R(Q(4))
 \end{aligned}$$

where  $R(Q(I))$  is the right hand side calculation. It should be noted that the Runge-Kutta algorithm has difficulty in converging to steady state if implemented as above. To accelerate convergence, implicit residual smoothing is applied (see [2]).

The current implementation of the PARC code also allows the user to separate a large problem into blocks, each of which can be solved separately. This allows for solution

of very large problems in systems with limited memory. After one or more time steps in a block, the boundary information is stored so that contiguous blocks may use it.

It should also be mentioned that the PARC code implements metrics to enable solutions to be calculated on general grids. As this has only minor impact on the algorithm we present here, we will not discuss it in this paper.

#### 4. Multigrid.

There are basically two types of multigrid methods which are implemented today. The first is used to accelerate the convergence of an implicit matrix solver operating on a system of difference equations used to approximate solutions to partial differential equations or systems of equations. The second is used on a time dependent problem to accelerate convergence to steady state. The PARC code can benefit from both types of algorithms. The quickest improvement, however, can be realized by implementing the second type of multigrid. This is due to the fact that there are no implicit matrix solver currently implemented in the PARC code.

In order to discuss the multigrid algorithm, we first need to establish some notation. Let  $M_1$  be the grid on which a solution is desired. Let  $M_i$  be the grid obtained by removing every other grid line from  $M_{i-1}$  in each coordinate direction. There is an  $n$ ,  $n \geq 1$ , such that  $M_n$  cannot be coarsened in the manner stated above. Also, let  $Q_i^n$

denote the calculated solution (by the solver of choice) on grid  $i$  at time step  $n$ .

Next, we define two operators which provide communication between solutions on

different grids, namely a prolongation operator  $I: M^i \rightarrow M^{i-1}$  and a restriction operator

$I^T: M^{i-1} \rightarrow M^i$ . For this discussion, the prolongation operators and restriction operators

can be viewed as interpolation and injection, respectively.

From the approximately factored algorithm generated from (2), we see that the system of equations to be solved on  $M^i$  can be written as

$$L_i(Q_i^n) \Delta Q_i^{n+1} = R(Q_i^n)$$

where  $L$  is the matrix constructed from the difference scheme.

The concept behind multigrid is to allow convergence to be computed in part by the coarse grid equations. As mesh size of the grid increases, the time step restriction given by the CFL condition can be relaxed. Thus, the coarse grid equations can march to the steady state condition faster than the fine grid equations. In order to retain the truncation error on the fine grid, the coarse grid solutions must be transmitted to the fine grid at regular intervals.

The algorithm currently implemented in the PARC code is due to Jameson [6]. The following is a description of the algorithm in pseudo code with the Beam Warming method as solver.

For level = 1 to number of levels do:

$$\text{Solve } L(Q_{\text{level}}^n) \Delta Q_{\text{level}}^{n+1} = R(Q_{\text{level}}^n) \text{ for } \Delta Q_{\text{level}}^{n+1}$$

$$\text{Update } Q_{level}^{n+1} \leftarrow Q_{level}^n + \Delta Q_{level}^n$$

If level < number of levels then do:

$$R_{level+1} \leftarrow I^T R_{level}(Q_{level}^{n+1})$$

$$Q_{level+1}^n = I^T Q_{level}^{n+1}$$

For level = number of levels - 1 to 2 do:

$$\Delta Q_{level}^n \leftarrow \Delta Q_{level}^n + I \Delta Q_{level+1}^n$$

$$Q_1^{n+1} \leftarrow Q_1^{n+1} + I \Delta Q_2^n$$

Hence, the coarse grids are given the current right hand side from the fine grids, and the old solution on the coarse grid is the new solution on the fine grid. After solutions are calculated on all grids, the differences are transferred to the fine grid.

The same general algorithm can also apply to the pseudo Runge-Kutta algorithm, but some modifications must be made. In the Runge-Kutta algorithm, R is calculated five times. Each time R must be modified by

$$R(Q(I)) \leftarrow R(Q(I)) + (I^T R(Q_{level-1}^{n+1}) - R(Q(0)))$$

In other words,  $R(Q(0))$  is merely a restriction of the fine grid right hand side. Further right hand side calculations are accounted for by the formula above.

### 5. Sample problem.

The problem we consider is flow through a two dimensional diverging nozzle with straight duct segments at each end. The initial conditions given the PARC code were appropriate for Mach .29 free stream flow. Reference pressure was at 15psia while the reference temperature was 600 degrees Rankine. The ratio of specific heats was 1.4. The inflow pressure and temperature agreed with the reference pressure and temperature, while the outflow boundary condition relied on a static pressure of 14.13psia. The domain and grid used to obtain a solution are given in Figures 1 and 2.

As the grid given in Figure 2 is 33 by 11, it can be coarsened only once and still retain agreement between coarse and fine grid points. For this reason, comparisons between the multigrid PARC code and the original PARC code were made on three grids. On the 33 by 11 grid, Multigrid operated with two levels. On the 65 by 21 grid it had 3 levels, while on the 129 by 41 grid there were 4 levels. The following table gives a comparison of the two methods on each grid. All solutions were due to calculations by the Beam-Warming solver as implemented in the PARC code. The number of iterations given in the table is the iteration count required for the L2 norm of the residual to be driven below  $10^{-12}$ .

Number of Levels	Number of Multigrid Iterations	Number of Regular Iterations	Cpu time for Multigrid solution	Cpu time for Regular solution
2	1550	2300	0.44	0.33
3	1550	3750	1.37	1.35
4	1500	6200	4.23	8.07

As can be seen above, the speed of convergence (number of iterations required) is approximately the same for each grid even though the grid sizes vary by a factor of two in each direction every time the grid is changed. The convergence seen is due to the contribution of the coarser grids. Since each problem given above has the same size grid for the coarsest mesh, the convergence remains roughly the same.

## 6. Conclusions.

The problem that we have attempted to address during the ten weeks of the 1991 Summer Faculty Fellowship program was implementation of multigrid and adaptive gridding techniques in the PARC code. This is, of course, an ambitious project which cannot be concluded in a ten week time frame. Therefore, the implementation of Multigrid in the PARC code was begun first. We have shown that there can be a significant improvement in the convergence of a fluid flow problem if Multigrid is applied with a large number of levels. Due to the extra overhead required for the Multigrid solver, a significant savings might not be realized with a small number of levels. This implies that extra care should be



made in choosing a grid which can be coarsened a large number of times. In many instances, a slight change in the mesh size in either direction would provide an appropriate grid.

Though the results given in this paper are encouraging, this certainly can not be viewed as the final implementation. There are many problems which still need to be addressed. Though the multigrid algorithm implemented in the PARC code at present only works for grids which have one block, it can be easily changed to work with multi-block problems. This can be accomplished by performing one Multigrid cycle on each block before moving to the next. A more efficient algorithm would cycle through each block on a level before going to the next level. An implementation of the second type, however, would require a different block interpolation routine than is currently available in the PARC code.

Self adaptive grids can be implemented an number of ways. In a finite difference code the easiest way to refine the grid would be to create a new block with the needed refinement. In order to solve on the new composite mesh, the boundary interpolation routines for the new block need to be changed to reflect conservation principles across the block boundary. Currently the PARC code does not implement this with sufficient accuracy.

Also, the PARC code is limited in the choice of available solvers. This should certainly be augmented to include a flux splitting solver with upwind differencing. This type of solver appears to do well on approximating supersonic flow. Other solvers could also be implemented if desired.

Finally, as was mentioned earlier, the boundary conditions of the PARC code are all explicit. Implicit boundary conditions could improve the performance of the Beam-Warming

solver or other implicit solvers. This could further increase the effectiveness of the PARC code.

The proposed changes to the PARC code mentioned above are intermediate steps to the goal mentioned at the beginning of this section. Implementation of adaptive refinement methods will also increase the effectiveness of the PARC code and lead to a production code which is able to find accurate solutions in less time.

## 7. References.

1. Anderson, D.A., J.C. Tannehill and R.H. Pletcher, Computational Fluid Mechanics and Heat Transfer, Hemisphere Publishing, New York, 1984.
2. Cooper, G. and J. Sirbaugh, "The PARC Distinction: A Practical Flow Simulator," AIAA paper 90-2002, 1990.
3. Dick, E., "A Multigrid Method for Steady Incompressible Navier-Stokes Equations and on Flux-Vector Splitting," Lecture Notes in Applied Mathematics 110, Marcel Dekker, New York, May, 1988.
4. Hanel, D., M. Meinke, and W. Schroder, "Application of the Multigrid Method in Solutions of the Compressible Navier-Stokes Equations," Proceedings of the Fourth Copper Mountain Conference on Multigrid Methods, SIAM, Philadelphia, 1989.
5. Huddleston, D.R., "Development of a Free-Jet Forebody Simulator Design Optimization Method," Arnold Engineering Development Center Technical Report AEDC-TR-90-22, Arnold Air Force Base, Tennessee.
6. Jameson, A., W. Schmidt, and E. Turkel, "Numerical Solutions of the Euler Equations

by Finite Volume Methods Using Runge-Kutta Time-Stepping Schemes," AIAA paper 81-1259, 1981.

7. McKay, S. and J.W. Thomas, "Application of the Self Adaptive Time Dependent Fast Adaptive Composite Grid Method," Proceedings of the Fourth Copper Mountain Conference on Multigrid Methods, SIAM, Philadelphia, 1989.
8. Siclari, M.J. and P. DelGuidice, "A Multigrid Finite Volume Method for Solving the Euler and Navier-Stokes Equations for High Speed Flows," AIAA paper 89-0283, 1989.
9. Volpe, G., Siclari, M.J., and Jameson, A., "A New Multigrid Euler Method for Fighter-Type Configurations," AIAA Paper 87-1160, presented at the AIAA 8th Computational Fluid Dynamics Conference, Honolulu, HI, June, 1987.
10. Mulder, W.A. and B. Van Leer, "Experiments with Implicit Upwind Methods for the Euler Equations," Journal of Computational Physics 59, pp. 232-246, 1985.

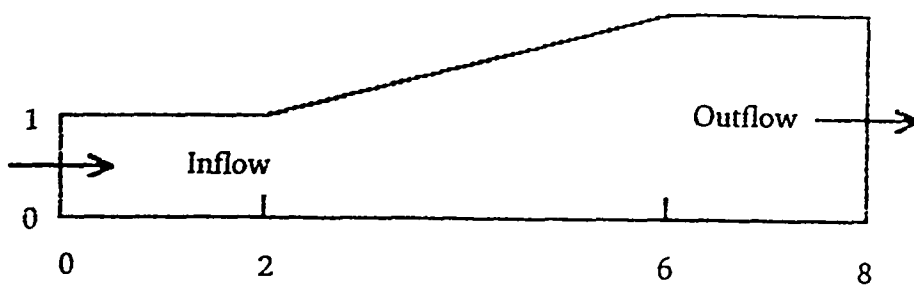


Figure 1. The domain of the flow problem presented in section 5.

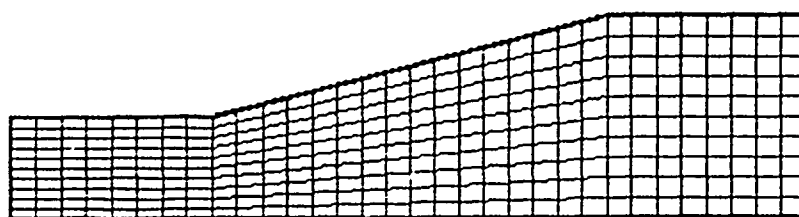


Figure 2. The 33 by 11 grid used for the solution of the problem given in section 5.

## X-RAY SPECTROMETERS FOR PULSED BREMSSTRAHLUNG

Carlyle E. Moore  
Morehouse College  
Atlanta, GA 30314

### ABSTRACT

The Photoactivation Spectrometer, Differential Absorption Spectrometer and Compton Spectrometer are examined as possible detection systems for the measurement of the spectra of pulsed bremsstrahlung. Detectors considered include the Ionization Chamber, PIN Junction Diode, Thermoluminescent Dosimeter (TLD) and Scintillator. The Photoactivation Spectrometer samples the spectrum at a relatively small number of points, and is limited by a lack of complete and reliable data on the nuclear parameters involved. Since the absorption characteristics of available absorbers do not vary significantly at energies above about 0.8 Mev, Differential Absorption Spectrometers show poor energy resolution in that region. They do, however, provide good results in the low energy end of the spectrum. The Time Projection Compton Spectrometer (TPCS) shows the greatest promise as a detection system for the measurement of pulsed X-ray spectra in the whole range of interest, 0.1 Mev to 2.0 Mev, although its deployment may be costly in terms of time, money and manpower.

### 1. INTRODUCTION

A proposal is under consideration to build a facility at AEDC for the purpose of testing space-bound military equipment in a nuclear-weapons environment, with a view to ensuring that it would survive a nuclear attack. The major threat comes from the release of X-radiation, which can result in a degradation of performance or even permanent damage. Simulation of nuclear-weapons radiation would be achieved by exposing the equipment to intense flash X-ray sources. These are generated by injecting high-energy pulsed electron beams into standard converter configurations, where they produce continuous X-rays (bremsstrahlung) together with characteristic line radiation. In order to implement this testing program, the energy spectrum of the radiation must be accurately known. The spectra emitted by pulsed X-ray simulators can be computed from the time histories of the diode current and voltage, based on the

predictions of a coupled electron-photon Monte Carlo code, such as the Integrated TIGER Series (ITS)<sup>1</sup>. However, the electron beam parameters are not well known, the photon spectrum depends on the angle of incidence of the electrons on the bremsstrahlung converter, and the simulators suffer from a lack of shot-to-shot reproducibility. It is therefore desirable to devise a technique for measuring the photon spectrum directly, with good precision, and on a shot-to-shot basis. The range of interest is from 0.1 Mev to 2.0 Mev.

For the purposes of this Report, an X-ray Spectrometer is defined as a device which is used to measure the spectrum of pulsed bremsstrahlung. The spectrum may be described either in terms of the spectral photon distribution  $U(e)$  or the spectral fluence  $F(e)$ , where  $U(e) de$  is the number of incident photons per unit area whose energy lies in the range  $e$  to  $e + de$ , and  $F(e) de$  is the energy per unit area of those photons. Clearly  $F(e) = e U(e)$ . An important characteristic of a radiation detector is its Response Function  $R(e)$ , which relates the input  $F(e)de$  to the detector to the measured output  $dM$  (charge collected, for example) as follows:

$$dM = R(e) F(e) de \quad (1.1)$$

$$M = \int R(e) F(e) de \quad (1.2)$$

A number of detection systems has been described recently in the literature<sup>2-8</sup>, using a variety of radiation detectors. Section II contains a discussion of the structure and use of these detectors, and in Section III we present an analysis of the detection systems themselves, including the techniques used to unfold the spectrum. In Section IV we offer some tentative conclusions about the relative merits of the systems considered.

## II. RADIATION DETECTORS

Broadly speaking, a radiation detector is simply a device which interacts with the incident radiation to produce some observable effect. It consists essentially of the following components:

1. A sensitive volume, containing the material which interacts with the radiation.
2. A supporting structure, which encloses the sensitive volume and provides the necessary instrumentation, if any. An essential part of this structure is the entrance window, through which the radiation is introduced into the sensitive volume.

3. An output mechanism, which extracts the output signal and presents it to the signal processing system.

A number of physical processes have been used as the basis for constructing practical radiation detectors, e.g. Nuclear Isomerism, Ionization, Thermoluminescence, etc., and each of these will be discussed in turn.

### Photoactivation Detectors

Excited nuclear states can be produced by a number of mechanisms, including electromagnetic excitation. The decay rate depends strongly on the angular momentum  $I$  of the excited state, and if there is a large difference in  $I$  between the excited state and all the lower states, transitions to these states are highly forbidden and the lifetime of the excited state turns out to be relatively long. These long-lived excited nuclear states are called Isomers. They have lifetimes ranging from a few seconds to several hours. Photoactivation of nuclear isomers proceeds via  $(\gamma, \gamma')$  reactions, in which a nucleus is excited from its ground state to a gateway, or activation, state by absorption of a photon  $\gamma$ , and subsequently decays to the metastable state (isomer) with emission of a photon  $\gamma'$ . The isomer then decays by process of delayed fluorescence, emitting a signature photon which may be used to identify the isomer. The isomer yield represents the output of the device, and is used to obtain information about the intensity of the incident radiation. The use of this technique has been discussed extensively in an earlier Report<sup>8</sup>, and will not be considered further here.

### Ionization Chambers

The ionizing effect of electromagnetic radiation is utilized in a number of devices to measure such quantities as the intensity of the radiation, the absorbed dose, etc. Extensive accounts of the design and operation of ionization chambers have been given by W. H. Tait<sup>9</sup> and by J. W. Boag<sup>10</sup>. The basic structure of the instrument is shown in Fig. 1. The material inside the sensitive volume is usually a gas and hence they are known generically as gas counters. They typically consist of two electrodes placed inside a chamber containing the gas. The electrodes may be co-axial cylinders or concentric spheres, but are more often a pair of parallel plates, separated by insulating spacers, and in what follows we shall assume that this is the case. The advantage of the parallel plate geometry is that a uniform electric field is maintained between the electrodes. However, this uniformity tends to break down near the edges of the plates, where the field lines curve outwards and the field strength gradually falls off to zero. To avoid this type of distortion, which makes it difficult to interpret the measured output of the detector, guard ring plates are placed around, and co-planar with, the electrodes. They are kept at the same potentials as the electrodes, but are not

connected to the signal processing system. The guard rings not only remove the non-uniformity of the field near the edges of the electrodes, they also help to define the sensitive volume more precisely.

The chamber is made of a solid which is rigid, impervious to the gas and free from chemical attack by it, and does not itself contaminate the gas by outgassing. Both the gas and the solid must be stable in the radiation environment. Ion pairs are created as a result of the interaction of the radiation with the gas. This interaction is a complex process<sup>12</sup>, in which there is a transfer of energy from photons to electrons via the photoelectric effect, Compton scattering and pair production, and electrons impart energy to the material by excitation, ionization and elastic scattering. The free electrons are attracted to the anode and the positive ions to the cathode. The charge collected by the electrodes may be allowed to accumulate and measured as a voltage pulse (charge mode) or allowed to leak away continuously in the form of an output current (current mode).

There is, of course, the question of recombination of electrons and positive ions to form neutral atoms, as well as the possibility that some electrons may acquire enough energy to cause further ionization. These depend on such factors as the potential difference  $V$  between the electrodes and the pressure of the gas. Appreciable recombination is expected to take place at small  $V$ , whereas secondary ionization occurs for large values of  $V$ , or at low gas pressures. We shall assume that the potential difference applied is large enough to avoid appreciable recombination, but not large enough to produce secondary ionization.

Since our ultimate objective is the measurement of the spectra of pulsed bremsstrahlung, we shall be concerned with the operation of the ionization chamber in charge mode. If, as is likely to be the case, the duration of the pulse ( $\approx$  ns) is much less than the time of transit of the ions across the ionization chamber, the ionization may be treated as taking place instantaneously. We shall assume, for the time being, that the ionization chamber is operating under ideal conditions. This implies that all the ion pairs formed in the chamber are collected by the applied field at the electrodes, without loss by recombination or augmentation by secondary ionization. The total charge (of either sign) collected is then given by

$$Q = n e = \frac{E}{W} e \quad (2.1)$$

where  $n$  is the number of ion pairs produced,  $e$  is the electronic charge,  $E$  is the absorbed energy and  $W$  is the mean energy required to produce an ion pair. A



summary of  $W$  values for a number of different gases, for several ionizing agents has been given by I. T. Myers<sup>13</sup>. As a result of the deposition of charge on the electrodes, a voltage pulse appears at the output end of the detector. This is illustrated in Fig. 2. Prior to the introduction of the radiation, a potential difference  $V$  is applied through a load resistance  $R$  across the detector plates, assumed to have a capacitance  $C$ . Charges  $q$  and  $-q$ , respectively, are deposited on the negative and positive plates, causing the potential difference across them to fall by an amount  $v = q/C$ . The capacitor is subsequently recharged by the power supply, with time constant  $RC$ . In practice, the signal pulse is extracted through a coupling capacitance  $C_c$ , which transmits only ac voltages, thereby isolating it from the much larger dc component. The output pulse then has the form shown in Fig. 2, rising linearly to its maximum height  $v$  and then decaying exponentially, with time constant  $RC$ .

The simple theory outlined above must be modified in light of certain practical considerations. The effect of recombination has been discussed extensively by Boag<sup>10</sup>. For pulsed radiation, it has been shown that the collection efficiency  $f$  (the ratio of charge collected to charge produced) is given by

$$f = \frac{1}{u_0} \ln(1+u_0); u_0 = \frac{\alpha}{k_1 + k_2} \left( \frac{d^2}{V} \rho_0 \right) \quad (2.2)$$

where  $\alpha$  is the recombination coefficient,  $k_1$  and  $k_2$  the mobilities of the positive and negative ions, respectively,  $d$  the separation between the plates and  $\rho_0$  the charge created per unit volume. For an accurate measurement of the total ionization produced, the chamber should be made air-equivalent, i.e. the chamber wall should have the same photon mass-energy absorption coefficient and the same mass stopping power as air<sup>9,10</sup>. This can be accomplished, approximately, by choosing the wall material to be a compound of low- $Z$  materials, such as nylon, carbon, silica (lined with graphite to provide adequate conduction), i.e. having an effective  $Z$  closely matching that of air. Practical considerations, such as ease of fabrication, may require the use of a wall material that is not air-equivalent, however. Tanaka et al<sup>11</sup> describe a parallel plate ionization chamber in which Aluminium is chosen as the wall material and the whole region between the plates is the sensitive volume.

### Solid State Detectors

The operation of a solid state detector is best understood in terms of the energy band structure of the material. The energy levels available to an electron in a crystal form a set of allowed bands which are separated by forbidden bands. In the ground state of the crystal, there may be several filled bands, the uppermost of which (the

valence band) is separated from the conduction band by an energy gap which is typically  $< 10$  eV for a solid insulator, but may be as little as 1 eV for a semiconductor. Since there are no vacant states in the allowed bands, there is no migration of charge within the crystal, unless electrons are given enough energy to surmount the forbidden energy gap and reach the conduction band. This energy can be supplied by, among other things, ionizing radiation. Electrons which reach the conduction band may migrate through the material under the action of an electric field. The elevation of an electron to the conduction band creates a vacancy or "hole" in the valence band which may be filled by a nearby valence electron, moving under the action of an electric field. The whole process is similar to the creation of an ion pair in a gas.

One of the essential features of an ionization chamber is that application of a moderate potential difference across the chamber should not result in the flow of current. The use of a solid insulator, in lieu of the gas in a conventional ion chamber, has two distinct advantages. First, the greater stopping power of a solid, as compared with a gas, should enhance the detection efficiency of the instrument. Secondly, the width of the energy gap between the valence band and the conduction band in a solid is typically much smaller than the ionization energy of a gas; hence the mean energy required to form an ion pair is much smaller in a solid than in a gas, and greater energy resolution should result.

The parallel plate ionization chamber with a solid insulator suffers, however, from a serious drawback, caused by the presence of impurities in the material. If the allowed energy states of these impurities are situated just below the conduction band of the host material, electrons normally resident in these states (donor states) are quite easily elevated into the conduction band by thermal excitation. This creates positive ions in fixed positions which act as "traps" for electrons from the signal pulse. On the other hand, discrete impurity levels may be located just above the valence band, and electrons from the valence band may be thermally excited into these impurity states (acceptor states), creating negative ions which can trap positive ions from the signal pulse. Crystal defects have a similar effect. These traps represent a serious departure from idealized ion chamber conditions, which can only be realized in an intrinsic material, i.e. one which is absolutely pure and perfect. Silicon and germanium are good candidates, but they are semiconductors, with energy gaps of the order of 1 eV between the valence and conduction bands, and therefore susceptible to leakage current at low applied potential differences or modest temperatures.

Now consider a single, high purity crystal (Si or Ge) which is doped on one side with n-type material and on the other with p-type material. The crystal will contain some free electrons in the conduction band and some free holes in the valence band. Thus the n-type material will contain many free electrons and very few holes, while the p-type material will contain many free holes but very few electrons. A current can be made to flow across the device if a potential difference is applied in such a direction as to attract the free charges toward the junction. This is known as forward bias. On the other hand, if the polarity of the applied potential is reversed (reverse bias), a region which is devoid of charge will be established in the vicinity of the junction. This region is known as the depletion region, and it behaves like an insulator, while the n- and p-type regions behave like conductors. The whole arrangement acts like an ionization chamber, with the depletion region being the sensitive volume. If ionizing radiation is introduced into the "chamber", it creates ion pairs (free electrons and holes) in the depletion region, and these charges migrate to the electrodes under the action of the applied potential. A Silicon p-n junction detector is shown schematically in Fig. 3. If all of these charges are collected (without trapping or recombination), the charge pulse is given by

$$Q = ne = \frac{E}{W}e \quad (2.1)$$

where  $n$  is the number of ion pairs formed,  $e$  is the electronic charge,  $W$  is the mean energy required per ion pair and  $E$  is the total energy deposited. It will be recalled that an identical expression was obtained for the gas ionization chamber.

The junction diode as outlined above suffers from the drawback, already alluded to, that electrons may be thermally excited across the narrow gap between the valence and conduction bands, producing a leakage current which increases with the applied (reverse bias) potential. Another source of leakage current is the presence of minority carriers, p-type impurities in the n-type material and vice versa. The thickness of the depletion region is proportional to  $\sqrt{\rho V}$ <sup>14</sup>, where  $\rho$  is the resistivity of Silicon and  $V$  is the applied voltage. This thickness determines the number of ion pairs formed, hence the detection efficiency of the device. To increase the thickness, either  $\rho$  or  $V$  must be increased. Increasing  $V$  would lead to insulation breakdown in the sensitive volume. The resistivity of Si can be effectively increased by diffusing carefully controlled amounts of  $\text{Li}_3$  ions into the lattice of a p-type crystal of Si, where they form neutral pairs with the negative acceptor ions. The increase in  $\rho$  means that larger bias voltages can be applied without causing electrical breakdown, and the thickness of

the depletion layer is thereby increased. Junction diodes with thick depletion layers produced in this way are called PIN junctions.

Consider a crystal insulator which is doped with a selected material which introduces impurity levels between the valence band and the conduction band. Irradiation of the crystal creates free electrons which are elevated to the conduction band, where they wander about in the solid until they either return to the valence band or are trapped at the impurity centers. The impurity acts like an activator, in the sense that the electrons remain trapped until they are subsequently released by some kind of excitation. If the release is caused by heating, thermoluminescence is said to occur. The trapped electrons are excited up to the conduction band, whence they return to the valence band with the emission of light. This process is shown schematically in Fig. 4. The amount of light emitted is proportional to the number of electrons released from the traps, which is in turn proportional to the absorbed dose. Thermoluminescence is a property of most solids, at least to some degree. To be useful as a radiation detector, however, a thermoluminescent phosphor must have a strong light output, and retain the trapped electrons for some time at the temperature at which it is to be used. Several materials exhibit these properties, including manganese activated Calcium Sulphate ( $\text{CaSO}_4:\text{Mn}$ ), manganese activated Calcium Fluoride ( $\text{CaF}_2:\text{Mn}$ ) and Lithium Fluoride. These phosphors are reviewed extensively by J. F. Fowler and F. H. Attix<sup>15</sup>. Instrumentation needed for readout of the TLD is also discussed.

### Scintillator Counters

By contrast, there are many solid and liquid substances which possess the property of instantaneous luminescence, or fluorescence. The incident radiation raises a number of electrons to excited energy states, from which they decay with the emission of flashes of light, or scintillations. Commonly used scintillators include organic and inorganic crystals, plastics, glasses, liquids and gases. The pulse of light emitted by a scintillator is typically very weak, and must be amplified by a photo-sensitive device, such as a photomultiplier. The light is emitted isotropically by the scintillator and must be channelled towards the photomultiplier by an optical coupling device, or light pipe. The scintillator, optical coupling and photomultiplier together constitute the radiation detection system, the scintillator being the sensitive volume. The whole assembly is enclosed in a light-tight container and the radiation is introduced through an entrance window. An extensive survey of scintillation detectors has been given by W. J. Ramm<sup>17</sup>.

### III. DETECTION SYSTEMS

#### Differential Absorption Spectrometers (DAS)

In designing an X-ray Spectrometer, the basic idea is to "sample" the spectrum at different points, using a number of different detectors. One of the limitations of the foil activation technique is that the spectrum can be sampled at relatively few points (these being determined by the activation levels of the nuclei involved) and there is only a small number of isotopes which are good candidates for this procedure. An alternative approach is to use a large number of identical detectors, and to interpose different thicknesses of different absorbing materials between them and the source of the radiation. Each detector then exhibits a different spectral response, and the spectrum can be unfolded once the detector response functions are known. The response of the  $i^{\text{th}}$  detector is, from Equation (1.2),

$$M_i = \int R_i(e) F(e) de \quad (3.1)$$

The index  $i$  runs from 1 to  $n$ , where  $n$  is the number of detectors. Equation (3.1) represents a system of  $n$  integral equations, which cannot be solved for arbitrary  $M_i$  and  $R_i(e)$ . To overcome this difficulty, the spectrum is divided into  $m$  energy "bins", of finite width, and equation (3.1) then becomes

$$M_i = \sum_{r=1}^m R_i(e_r) F(e_r) \Delta e_r \quad (3.2)$$

where  $\Delta e_r$  is the width of the  $r^{\text{th}}$  energy bin.

The response functions may be calculated<sup>4</sup> if the mass energy absorption coefficients of the absorbers and the detector material are known. If an absorber of thickness  $x_i$  and density  $\rho_i$  is interposed between the source and the  $i^{\text{th}}$  detector, the fluence  $F(e)$  of the incident radiation is attenuated by a factor

$$\exp\left[-(\mu_{en}/\rho)_i(e)\rho_i x_i\right]. \quad \text{Subsequently, a fraction } \left[1 - \exp\left\{-\left(\frac{\mu_{en}}{\rho}\right)_D \rho_D x_D\right\}\right] \text{ of the}$$

energy striking the detector is absorbed by it, where the subscript  $D$  refers to the detector. Thus, assuming that the measured output of the detector is proportional to the absorbed energy, we have

$$dM_i = \text{constant} \cdot \exp\left[-(\mu_{en}/\rho)_i(e)\rho_i x_i\right] \left[1 - \exp\left\{-\left(\frac{\mu_{en}}{\rho}\right)_D \rho_D x_D\right\}\right] F(e) \Delta e \quad (3.3)$$

whence the response function can be found by comparison with equation (1.1). It is assumed in the above discussion<sup>4</sup> that there is electronic equilibrium between the absorber and the detector (i.e. their atomic numbers are approximately equal), so that the loss of electrons which leave the detector without being absorbed is balanced by an equal gain of electrons entering the detector from the absorber. Where electronic equilibrium does not exist, the response functions must be found from a full Monte Carlo radiation transport computation. The transport codes TIGER<sup>1</sup> or the discrete ordinates code CEPXS-ONETRAN<sup>18</sup> can be used for this purpose.

A Differential Absorption Spectrometer must be used in conjunction with an unfolding code or technique which allows the spectral fluence  $F(e)$  to be calculated from Equation (3.1). A computer code named YOGI<sup>19</sup> has been developed by T. L. Johnson and S. G. Gorbics and used in the unfolding process. A trial spectrum and the known response functions are input into the program, which yields calculated values of the detector response. These are compared with the measured values  $M_i$  and a root mean square deviation is determined. The code then perturbs the trial spectrum at some point by a fixed amount and a new rms deviation is found. The perturbed spectrum is retained if it yields a better fit (smaller rms error), otherwise it is discarded. The procedure is repeated for randomly chosen points until some predetermined level of accuracy is reached or no further reductions in error can be achieved. It should be noted that Equation (3.1) represents a set of  $n$  equations in  $m$  unknowns. Thus if  $n$  is less than  $m$ , the problem is underdetermined and an infinite number of solutions is possible. In such a case, constraints must be imposed at each perturbation in the unfolding routine, to ensure that only physically acceptable solutions are considered. A "realistic" trial spectrum must be used, negative values should be rejected and the solution should be continuous except for the K-line in a specific energy bin.

A differential absorption spectrometer has been developed and used by G. A. Carlson and L. J. Lorence<sup>3</sup> on various bremsstrahlung sources. The arrangement utilized a stack of thirteen flat metal absorbers (2 of Al, 5 of Cu and 6 of a 90% Tungsten alloy), separated by TLD arrays. Each array consisted of four  $\text{CaF}_2\text{:Mn}$  TLD's placed in an Aluminum tray which was inserted into the DAS. The DAS was fielded at a distance of at least 30 cm from the source, without collimation or shielding. The spectrum was divided into thirty energy bins, equally spaced on a logarithmic scale. The unfolded spectra were found to be in reasonable agreement with TIGERP predictions.

A DAS which incorporates a slightly different design has been developed by S. G. Gorbics and N. R. Pereira<sup>20</sup>. This uses an array of spherical absorbers, each with a TLD embedded at its centre. This has the advantage that the absorption distance is accurately known (the radius of the sphere), irrespective of the direction of the incident radiation.

### Compton Spectrometers

As mentioned previously, when X-radiation interacts with matter, it may impart energy to electrons via the photoelectric effect, Compton scattering and pair production. The relative importance of these mechanisms<sup>16</sup>, for various atomic numbers  $Z$  and photon energies, is shown in Fig. 5. For low  $Z$  materials and incident photon energies in the range of interest (100 keV to 2.0 MeV), Compton scattering is clearly the dominant process. Pair production cannot occur at all if the photon energy is less than the rest energy of the electron-positron pair (1.02 MeV) and is not significant in low  $Z$  materials until much higher energy. Photoemission can also be ruled out if the detector is designed to reject low-energy electrons. Compton scattering thus becomes an attractive mechanism for the design of an X-ray spectrometer. A collimated beam of radiation is passed through a magnetic field to remove any electrons which may be present and then impinges on a thin, low  $Z$  scattering foil. The extraction of the beam of recoil electrons and measurement of the signal can be performed in either of two ways.

The magnetic Compton Spectrometer is a device in which the electrons scattered in some pre-determined direction are collimated and their energy determined by using the Klein-Nishina cross-section formula for Compton scattering and making corrections for deviations in the path of the electron and loss of energy due to multiple scattering as it passes through the foil. This technique has the disadvantage of being very inefficient, however. Only a narrowly collimated fraction of the electron emission is meant to be detected and extensive shielding is required. Nonetheless, this technique has been used to measure the spectrum of a 10 MeV flash X-ray generator<sup>21</sup>. The **Time Projection Compton Spectrometer (TPCS)**<sup>5-7</sup>, developed at SANDIA, is an alternative device in which the electrons emitted from the converter foil migrate towards the detector in a narrow, evacuated drift tube with an axial current-carrying wire. The magnetic field created is inversely proportional to the distance from the wire and the electrons experience a net drift along the axis (in the opposite direction to the current) with a drift velocity (see Appendix ) given by

$$v_d = \frac{1}{2} \epsilon (1 + \cos^2 \theta) \gamma \quad (3.4)$$

where

$$\epsilon = \frac{2\pi \gamma m v}{\mu_0 I e} \quad (3.5)$$

$v$  is the (constant) speed of the electron,  $I$  is the current,  $e$  and  $m$  are the charge and mass, respectively, of the electron,  $\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}}$  and  $\theta$  is the angle which the initial velocity of the electron makes with the magnetic field.

A schematic diagram of the TPCS is shown in Fig. 6. The foil is fixed in a radial plane of the drift tube and is irradiated at normal incidence by a collimated beam of X-rays. In order to maximize the signal, the thickness and the cross-section of the foil should be made as large as possible. However, the extent to which this can be done is limited by certain practical considerations. Multiple scattering within the foil changes both the direction and energy of the electrons, and therefore distorts the energy spectrum. This places a limit on the thickness of the foil. The axial dimension of the foil must be less than the pitch of the helical orbit, otherwise the electron would collide with the foil after the first orbit. The radial dimension of the foil should be equal to that of the detector (a plastic scintillator, coupled to a fast photomultiplier tube).

The TPCS differs from the DAS in that there is only one detector, and that the required photon spectrum is mapped into a time domain by the time of flight measurements. Thus the response function  $R(e, t)$  is now a function of both energy and time. The signal  $S(t)$  measured by the detector at time  $t$  is given by

$$S(t) = \int R(e, t) Y(e) de \quad (3.6)$$

where  $Y(e)de$  is the number of photons whose energies lie in the range  $e$  to  $e + de$ . In practice, the energy spectrum is divided into a number of energy bins of finite width  $\Delta e_j$ , so that Equation (3.6) is replaced by

$$S(t) = \sum_{j=1}^m R_j(t) Y(e_j) \Delta e_j \quad (3.7)$$

where  $m$  is the number of energy bins. The response function  $R(e, t)$  is computed by using a Monte Carlo simulation of the X-ray interaction. The number of photons per unit area with energy  $e$  which strike the foil and are scattered through the (polar) angle



$\theta$  ( $\cos\theta = \mu$ ) is given in terms of the Klein-Nishina differential cross-section  $d\sigma(\epsilon, \mu)$  by

$$dN = U(\epsilon) n_e T d\sigma(\epsilon, \mu) d\epsilon \quad (3.8)$$

where  $n_e$  is the number of electrons per unit volume and  $T$  is the thickness of the foil. The kinetic energy and direction of the recoil electron are obtained by applying conservation of energy and momentum. The azimuthal angles of the electron direction are assumed to be distributed uniformly between 0 and  $2\pi$ . We thus obtain the number of electrons which are emitted with a given speed  $v$  in a given direction. From Equation (3.4), which gives the drift velocity of the electrons, the number of electrons with transit time  $t$  produced by X-rays of energy  $\epsilon$  is computed. This number is clearly proportional to  $dN$  (see Equation (3.8)), and hence the measured signal satisfies the relation

$$dS(\epsilon) \propto n_e T d\sigma(\epsilon, \mu) U(\epsilon) d\epsilon \quad (3.9)$$

or

$$dS(\epsilon) = R(\epsilon) U(\epsilon) d\epsilon \quad (3.10)$$

where  $R(\epsilon)$  is the response function.

The unfolding of the spectrum, from Equation (3.7), consists of a least squares fit to the measured signal  $S(t)$ , using a trial spectrum  $N'(e_k)\Delta e_k$ . A variational principle is applied which minimizes the square of the difference between the measured signal  $S(t)$  and the signal calculated by using the trial spectrum. This yields<sup>6</sup>

$$\int S(\epsilon) R_j(\epsilon) d\epsilon = \sum_{j=1}^m U(\epsilon_k) \Delta e_k \cdot \int R_j(\epsilon) R_k(\epsilon) d\epsilon \quad (3.11)$$

which may be solved by matrix methods to determine the unknown spectrum  $U(\epsilon)$ .

#### IV. DISCUSSION

We have examined a number of X-ray Spectrometers, which utilize a variety of radiation detectors and unfolding techniques. The photoactivation spectrometer is limited by the fact that it samples the X-ray spectrum only at the energies of the activation levels and that there is a relatively small number of nuclei which are good candidates for this procedure. It also suffers from a lack of complete and reliable data on the nuclear parameters of interest. Of the detectors considered, the TLD is perhaps the most popular because it is small, inexpensive, available in a wide range of materials and dose ranges, requires no instrumentation during irradiation and retains accurate

dose information for a long time. The ionization chamber, by contrast, would appear to require extensive testing to determine the optimum choices of its size and shape, electrode material, type of gas and pressure, etc. The absorbers used in the Differential Absorption Spectrometer do not show sufficient variation in their absorption characteristics to provide adequate energy resolution above about 800 kev, and their use is thus effectively limited to the low energy region. The Time Projection Compton Spectrometer, on the other hand, can be used to effect throughout the region of interest (0.1 Mev to 2.0 Mev). It is, however, cumbersome and expensive to field and practical considerations may well restrict its use to the high energy end of the X-ray spectrum.

### APPENDIX

The following is adapted from a Report by J. R. Lee<sup>22</sup> of Sandia National Laboratories. Consider the motion of a relativistic electron in the magnetic field due to a current  $I$  flowing in a long straight wire in the negative  $z$ -direction. In a cylindrical coordinate system  $(s, \phi, z)$  the magnetic field is given by

$$\vec{B} = -\frac{\mu_0 I}{2\pi s} \hat{\phi} \quad (\text{A.1})$$

and the equation of motion of the electron is

$$\gamma m \frac{d\vec{v}}{dt} = \vec{F} - (\vec{F} \cdot \vec{v}) \frac{\vec{v}}{c^2} \quad (\text{A.2})$$

where

$$\vec{F} = -e(\vec{v} \times \vec{B}) = \left( \frac{\mu_0}{2\pi s} I e \right) (\vec{v} \times \hat{\phi}) \quad (\text{A.3})$$

Since the magnetic field does no work on the electron ( $\vec{F} \cdot \vec{v} = 0$ ), the speed  $v$  is constant and the equation of motion becomes

$$\frac{d\vec{v}}{dt} = \left( \frac{\mu_0}{2\pi s} \frac{I e}{\gamma m} \right) (\vec{v} \times \hat{\phi}) \quad (\text{A.4})$$

The constant  $\frac{\mu_0}{2\pi} \frac{I e}{\gamma m}$  has the dimensions of velocity, and it will be convenient to introduce the dimensionless constant

$$\varepsilon = \frac{2\pi \gamma m v}{\mu_0 I e} \quad (\text{A.5})$$

and write Equation (A.4) in the form

$$\frac{d\vec{v}}{dt} = \frac{1}{\epsilon s} (\vec{v} \times \hat{\phi}) \quad (\text{A.6})$$

The components  $v_s$ ,  $v_\phi$  and  $v_z$  satisfy the equations

$$v_s - \frac{1}{s} v_\phi^2 = -\frac{v}{\epsilon s} v_z \quad (\text{A.7a})$$

$$v_\phi + \frac{1}{s} v v_s = 0 \quad (\text{A.7b})$$

$$v_z = \frac{v}{\epsilon s} v_s \quad (\text{A.7c})$$

Combining Equations (A.7b) and (A.7c), we see that  $\frac{d}{dt} \left\{ v_\phi \exp \left[ \frac{\epsilon}{v} v_z \right] \right\} = 0$ , i.e.

$$\xi = \frac{v_\phi}{v} \exp \left[ \frac{\epsilon}{v} v_z \right] \quad (\text{A.8})$$

is a constant of the motion. Similarly, it follows from Equation (A.7c) that

$$s_0 = s \exp \left[ -\frac{\epsilon}{v} v_z \right] \quad (\text{A.9})$$

is a constant of the motion. Multiplying (A.8) and (A.9), we see that  $sv_\phi$  (and hence  $s \cos \theta$ ) is also a constant of the motion.

The decomposition of the velocity vector  $\vec{v}$  into its components  $v_s$ ,  $v_\phi$  and  $v_z$  is shown in Fig. 7, where

$$v_\phi = v \cos \theta \quad (\text{A.10a})$$

$$v_s = v \sin \theta \sin \alpha \quad (\text{A.10b})$$

$$v_z = v \sin \theta \cos \alpha \quad (\text{A.10c})$$

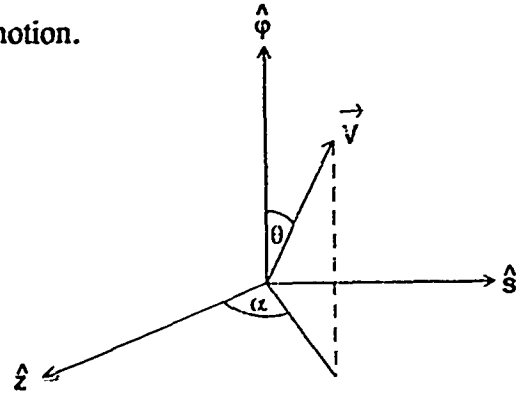


Fig. 7. Decomposition of electron velocity

Substituting Equations (A.10) into Equations

(A.7), we obtain a pair of coupled equations in the two unknowns  $\theta$  and  $\alpha$ :

$$\dot{\theta} = \frac{v}{s} \sin \alpha \cos \theta \quad (\text{A.11a})$$

$$\dot{\alpha} = -\frac{v}{\epsilon s} [1 - \epsilon \cos \alpha \cos \theta \cot \theta] \quad (\text{A.11b})$$

The drift velocity  $v_d$  of the electron is taken to be the average of  $v_z$  over one orbit ( $2\pi$  increment in  $\alpha$ ). We may thus write  $v_d = Z/T$ , where

$$Z = \epsilon s_0 \int_0^{2\pi} \frac{\sin\theta \cos\alpha \exp[\epsilon \sin\theta \cos\alpha]}{1 - \cos\alpha \cos\theta \cot\theta} d\alpha ; T = \frac{\epsilon s_0}{v} \int_0^{2\pi} \frac{\exp[\epsilon \sin\theta \cos\alpha]}{1 - \cos\alpha \cos\theta \cot\theta} d\alpha \quad (\text{A.12})$$

After a lengthy calculation, power series solutions in  $\epsilon$  for  $Z$  and  $T$  are obtained. To first order in  $\epsilon$ , this gives

$$v_d = \frac{1}{2} \epsilon v (1 + \xi^2) \quad (\text{A.13})$$

It is estimated<sup>22</sup> that for  $\epsilon = 0.05$ , the maximum error in computing the drift velocity  $v_d$  is less than 1%.

### Acknowledgements

I would like to thank the Air Force Systems Command, Air Force Office of Scientific Research, for their sponsorship of this research and the Arnold Engineering Development Center, Arnold AFB, TN for its hospitality during the summer. I also wish to thank Lavell Whitehead and Tim Cotter of CALSPAN for suggesting this topic and for their support throughout my stay. I owe a special debt of gratitude to Sid Steely for his kindness, patience and inspiration.

### REFERENCES

1. J. A. Halbleib and T. A. Melhorn, Nucl. Sci. Eng. 92, No.2, 338 (1986).
2. S. G. Prussin, S. M. Lane, D. R. Kania, J. E. Trebes and M. Krishnan, IR&D Project #50020921, conducted at the Physics International Company.
3. G. A. Carlson and L. J. Lorence, IEEE Trans. Nucl. Sci., NS-35, No. 6, 1255 (1988).
4. N. R. Pereira and S. G. Gorbics, private communication.
5. G. T. Baldwin and J. R. Lee, IEEE Trans. Nucl. Sci., NS-33, 6, 1298 (1986).
6. G. T. Baldwin, C. O. Landron, J. R. Lee, R. J. Leeper and L. J. Lorence, Jr., private communication.
7. G. T. Baldwin, C. O. Landron, J. R. Lee, R. J. Leeper and L. J. Lorence, Jr., Nuclear Instruments and Methods in Physics Research A299, 66 (1990).
8. C. E. Moore, AFOSR Report, Contract No. F49620-88-C-0053 (1990).
9. W. H. Tait, Radiation Detection, (Butterworths, 1980)
10. J. W. Boag, in Radiation Dosimetry, Vol. II (Second Edition), edited by F. H. Attix and W. C. Roesch (Academic Press, 1966).

11. R. Tanaka, H. Kaneko, N. Tamura, A. Katoh and Y. Moriuchi, Proceedings of the IAEA symposium SM-272/17, Vienna (1984).
12. J. H. Hubbell, NSRDS-NBS Handbook 29, National Bureau of Standards, Washington (1969).
13. I. T. Myers, in Radiation Dosimetry, Vol. 1 (Second Edition), edited by F. H. Attix and W. C. Roesch (Academic Press, 1969), p 320.
14. J. W. Fowler in Radiation Dosimetry, Vol. II (Second Edition), edited by F. H. Attix and W. C. Roesch (Academic Press, 1966), p295.
15. J. W. Fowler and F. H. Attix, in Radiation Dosimetry, Vol. II (Second Edition), edited by F. H. Attix and W. C. Roesch (Academic Press, 1966), pp272 ff.
16. R. D. Evans, in Radiation Dosimetry, Vol. 1 (Second Edition), edited by F. H. Attix and W. C. Roesch (Academic Press, 1969), p 97.
17. W. J. Ramm, in Radiation Dosimetry, Vol. II (Second Edition), edited by F. H. Attix and W. C. Roesch (Academic Press, 1966).
18. L. J. Lorence, Jr., W. E. Nelson and J. E. Morel, IEEE Trans. Nucl. Sci., NS-32, No. 6, 4416, (1981).
19. T. L. Johnson and S. G. Gorbics, Health Physics, 41, 859 (1981).
20. S. G. Gorbics and N. R. Pereira, J. of Rad. Effects, 6 (I), 64 (1988).
21. J. G. Kelley, L. D. Posey and J. A. Halbleib, IEEE Trans. Nucl. Sci., NS-18, No. 2, 131 (1971).
22. J. R. Lee, Sandia National Laboratories Report SAND89-0241 (March, 1989).

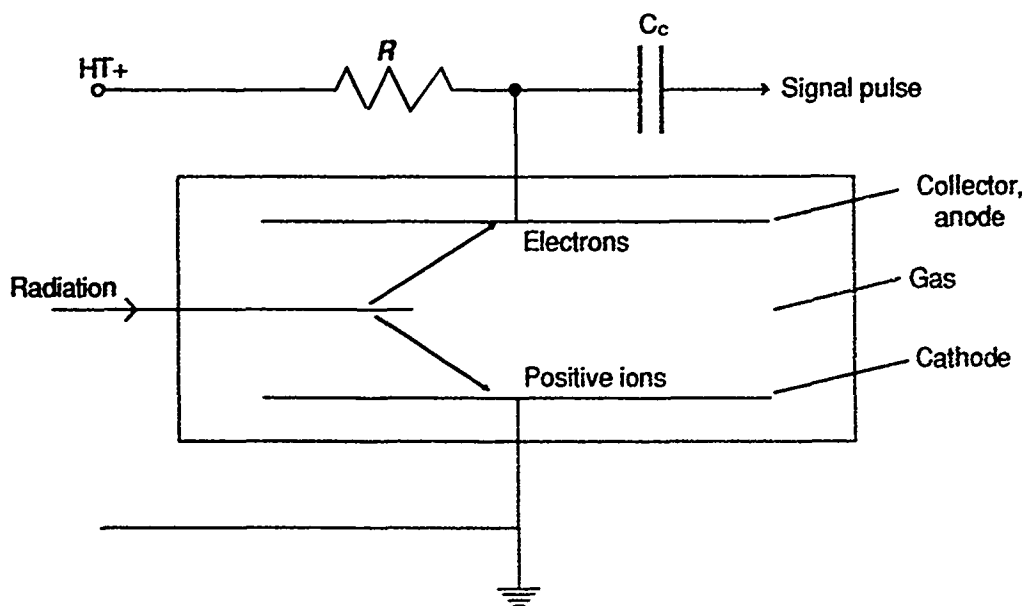


Fig. 1. Basic structure of an ionization chamber (from Ref. 9).

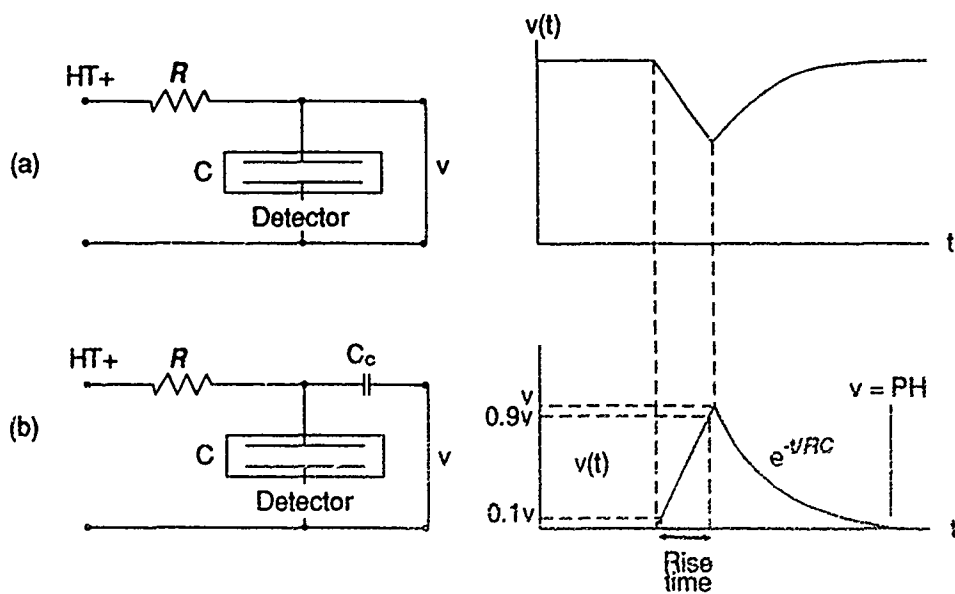


Fig. 2. Output pulse from detector circuit.  
(a) d.c. coupled (b) a.c. coupled (from Ref. 9).

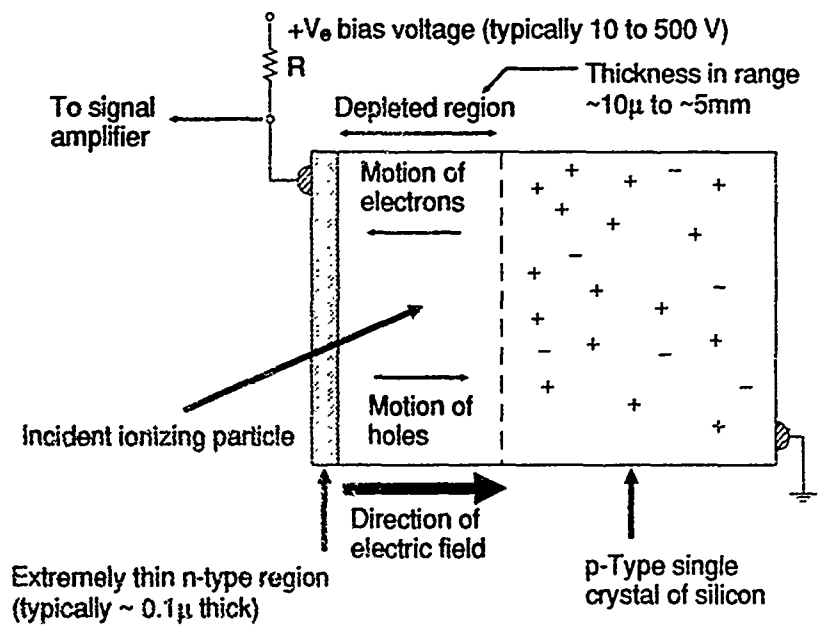


Fig. 3. Silicon p-n junction radiation detector (from Ref. 14).

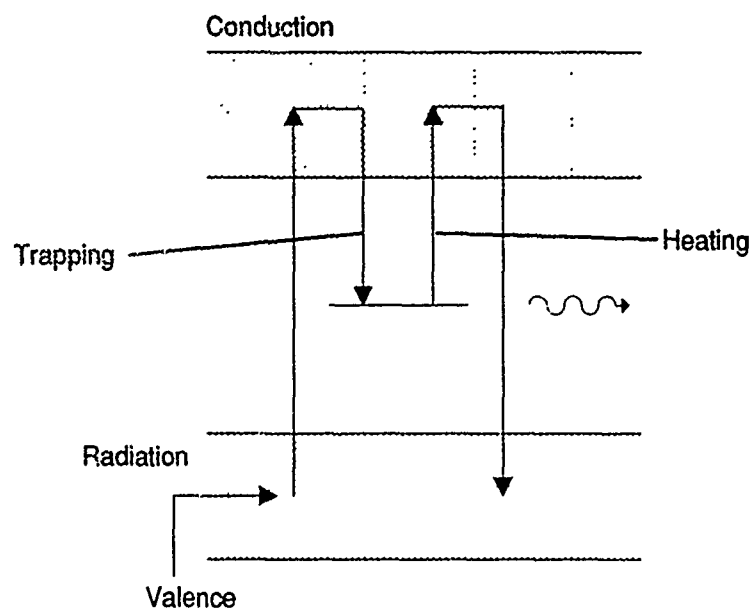


Fig. 4. The thermoluminescence process (from Ref. 9).

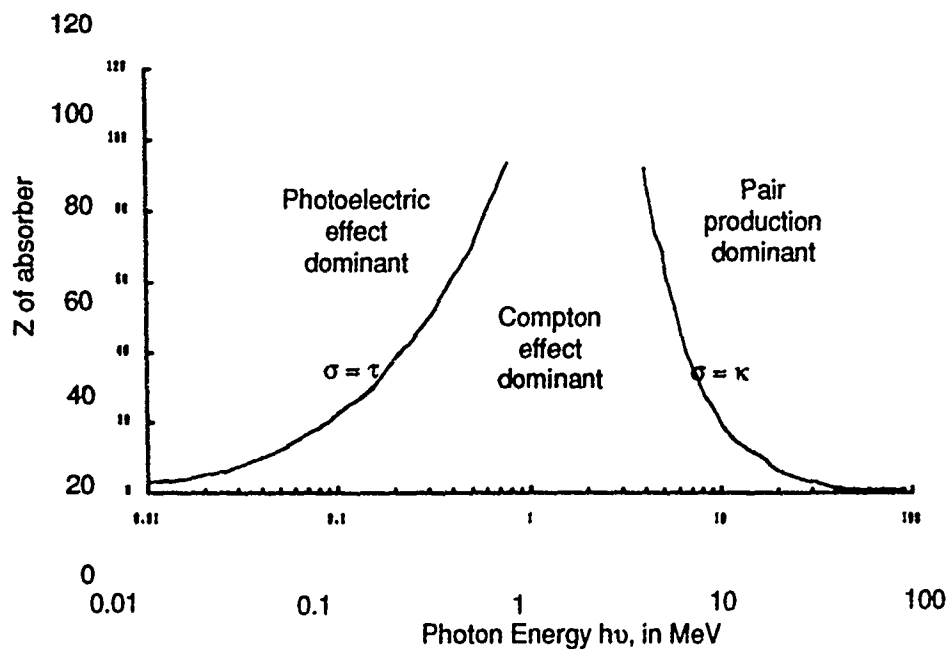


Fig. 5. Relative importance of x-ray interactions (from Ref. 16).

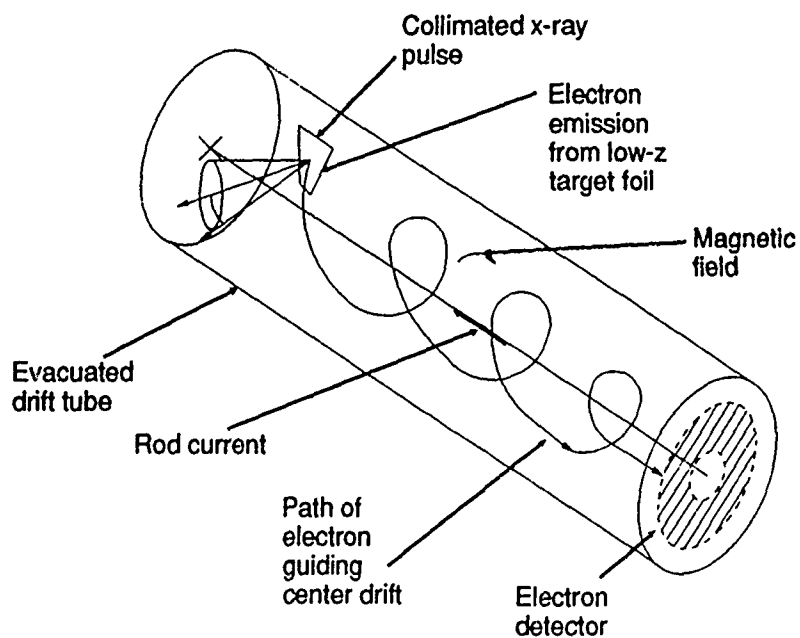


Fig. 6. Schematic representation of the Time Projection Compton Spectrometer (from ref. 6)



# THE EFFECT OF CARBON PARTICLE COMBUSTION ON THE INFRARED SIGNATURE OF A MAGNESIUM-FLUOROCARBON FLARE

Dr. Olin Perry Norton

## Abstract

A burning magnesium-Teflon flare produces a plume which contains a substantial quantity of solid carbon particles. As the flare products mix with the surrounding air, these carbon particles will burn. The combustion of carbon particles has been incorporated into SMIRF (Signature Model for InfraRed Flares), an existing model for predicting the plume structure and signature of these flares. It is shown that the inclusion of carbon particle combustion has a significant effect on the predicted plume signatures.

## Introduction

Teflon-magnesium flares are used as infrared decoys. Thus, the accurate prediction of their infrared signatures is of considerable interest. A computer model for the prediction of flare signatures, SMIRF (Signature Model for InfraRed Flares), has been developed at Arnold Engineering Development Center.<sup>1</sup> This code uses the existing JANNAF codes, originally developed for rocket plume signature prediction, and modifies them to solve the related problem of predicting the flowfield and infrared signature in the plume (or wake) of a Teflon-magnesium flare.

Throughout this report, results will be cited for a "typical" flare. This is a hypothetical flare whose characteristics are representative of actual flares. For this report, this generic flare is a rectangular parallelepiped which measures 2x3x5 inches and has a mass of 1.4 lbm. The flare composition is 54% (by mass) Mg, 30% C<sub>2</sub>F<sub>4</sub>, and 16% C<sub>2.21</sub>F<sub>2.84</sub>H<sub>1.58</sub>. This solid burns at a rate of 0.144 inches/second, giving the flare a lifetime of 6.9 seconds from ignition to burnout. For the calculations to be presented in this paper, the flare is ignited and ejected from an aircraft at an altitude of 9097 feet, and has an initial velocity of 422 ft/sec horizontally and -15 ft/sec vertically.

This flare is extremely fuel rich. (The equivalence ratio is 5.) When it burns, the resulting combustion products contain several species that will undergo further combustion when mixed with air. The NASA chemical equilibrium code<sup>2</sup> has been used to predict the composition of the combustion products of this typical flare. The results are given in Table 1.

Table 1. Flare Combustion Products		
Composition	Mg(gas)	38.8% mole fraction
	C(graphite)	30.2%
	MgF <sub>2</sub> (liquid)	24.2%
	H <sub>2</sub> (gas)	4.5%
	MgF(gas)	1.5%
	Remainder	< 1%
Temperature	1913 K	

Thus, the products from the combustion of the solid flare material contain significant quantities of magnesium vapor and solid carbon (graphite) particles, both of which will burn as the plume mixes with the surrounding atmosphere. Further calculations were made with the equilibrium code which demonstrated that mixtures of flare products with air will reach temperatures above 2900 K. This is a significant increase above the 1913 K temperature of the flare product itself and demonstrates the potential influence of afterburning on the temperature distribution.

The solid carbon particles will be significant emitters of infrared radiation, so the accurate prediction of the distribution of these particles will be important in predicting the plume signature. Furthermore, the combustion of the carbon particles is the only possible source of CO and CO<sub>2</sub>, gases which are important emitters of infrared information.

Based on these considerations, an effort was begun to accurately model the combustion of carbon particles in the plume of a Teflon-magnesium flare. In previous work, a literature survey was undertaken to identify the species in the plume which react with solid carbon.<sup>3,4</sup> These species were identified as the molecules O<sub>2</sub>, CO<sub>2</sub>, and H<sub>2</sub>O, and the radicals OH, O, and H. Rates were found for the reactions of these species with solid carbon surfaces. These rates are reprinted in Table 2.

Table 2. Kinetic Rates for Reactivity of Carbon Surface with Gases	
O <sub>2</sub>	$\omega_{O_2 \text{ chem}} / (12 \text{ g/mole}) = (\kappa_A P_{O_2} / (1 + \kappa_Z P_{O_2})) \chi + \kappa_B P_{O_2} (1 - \chi)$ where $\chi = (1 + \kappa_T / (\kappa_B P_{O_2}))^{-1}$ $\kappa_A = (20 \text{ moles}/(\text{cm}^2 \cdot \text{sec} \cdot \text{atm})) \exp(-15100/T)$ $\kappa_B = (4.46 \times 10^{-3} \text{ moles}/(\text{cm}^2 \cdot \text{sec} \cdot \text{atm})) \exp(-7640/T)$ $\kappa_T = (1.51 \times 10^{-5} \text{ moles}/(\text{cm}^2 \cdot \text{sec})) \exp(-48800/T)$ $\kappa_Z = (21.3 \text{ atm}^{-1}) \exp(2060/T)$
CO <sub>2</sub>	$\omega_{CO_2 \text{ chem}} = (247 \text{ g}/(\text{cm}^2 \cdot \text{sec} \cdot \text{atm})) P_{CO_2} \exp(-21100/T)$
H <sub>2</sub> O	$\omega_{H_2O \text{ chem}} = (247 \text{ g}/(\text{cm}^2 \cdot \text{sec} \cdot \text{atm})) P_{H_2O} \exp(-21100/T)$
OH	$\omega_{OH \text{ chem}} = (0.28) (12 \text{ g/mole}) \bar{\rho}_{OH} \sqrt{KT/2\pi(17 \text{ g/mole})}$
O	$\omega_{O \text{ chem}} = (0.50) (12 \text{ g/mole}) \bar{\rho}_O \sqrt{KT/2\pi(16 \text{ g/mole})}$
H	$\omega_{H \text{ chem}} = (0.036) (12 \text{ g/mole}) \bar{\rho}_H \sqrt{KT/2\pi(1 \text{ g/mole})}$
<p><math>\omega_{X \text{ chem}}</math> is the chemical reaction rate at which species X removes carbon from the surface in units of grams per second per square centimeter of surface.</p> <p><math>P_X</math> is the partial pressure of species X in atmospheres.</p> <p><math>T</math> is the surface temperature in degrees Kelvin.</p> <p><math>\bar{\rho}_X</math> denotes the concentration of species X in moles per unit volume.</p> <p><math>K</math> is Boltzmann's constant.</p>	

When these rate expressions were evaluated under typical plume conditions the species  $\text{CO}_2$ , OH, and O were the most important reactants with the solid carbon. It was found that, for particles less than 10 microns in diameter, the particle combustion was kinetically rather than diffusion limited, so that the reactant concentration at the particle surface could be approximated by its bulk concentration. Furthermore, any temperature difference between the gas and the particles decreases as the particle size is reduced, so that small particles can be regarded as being at the same temperature as the gas.

References 3 and 4 contain more complete descriptions of the previous work on this problem, including bibliographical references for the reaction rates in Table 2. The present work is a continuation of this previous work on carbon particle combustion, and consists of incorporating these carbon particle reactions into the SMIRF flare model and demonstrating their effects on the predicted signature.

### Discussion

The objective is to include reactions for solid carbon particles in the existing SMIRF flare model. To calculate the structure of a reacting plume, SMIRF includes a modified version of SPF, the standard JANNAF plume model.<sup>5</sup> Thus, the problem becomes one of including a rudimentary description of reacting particles in SPF.

The difficulty posed is this: although SPF does have the ability to include some particle effects, this ability is restricted to nonreacting particles of aluminum oxide. The solid species carbon will be treated as if it were a gas, in that SPF will compute a "concentration" of solid carbon (so many moles of solid carbon per unit volume) at each grid point. This pseudo-gas approximation creates difficulties when calculating chemical reaction rates involving the particulate species. The reaction types included in SPF are valid for chemically reacting gaseous species. For example, for a reaction between two gaseous species called A and B, the reaction rate is written in terms of the concentrations of the two species:

$$k \bar{\rho}_A \bar{\rho}_B \quad (1)$$

where  $k$  is the rate constant, which may be a function of temperature, and  $\bar{\rho}_A$  and  $\bar{\rho}_B$  represent the concentrations of species A and B. Depending on the units chosen for  $k$ , these concentrations are usually expressed as moles per cubic centimeter or grams per cubic centimeter.

The reaction rate should be calculated differently if one species (B for example) is in the form of a particle or droplet. Assuming that the reactions are kinetically limited, the rate of the reaction is proportional to the exposed surface area of species B:

$$k' \bar{\rho}_A \bar{\sigma}_B \quad (2)$$

where  $\bar{\sigma}_B$  is the total exposed surface area of the particles of species B per unit volume and  $k'$  is the surface rate constant.

#### Uniform Particle Diameter Approximation

The simplest model for the particle size distribution of the solid carbon particles is to assume that all particles are spheres of uniform diameter. Letting  $\eta$  be the number of carbon particles per unit volume and  $d$  stand for their diameter, then the pseudo-gas concentration of carbon is

$$\bar{\rho}_c = \frac{\pi}{6} \rho_c \eta d^3 \quad (3)$$

where  $\rho_c$  is the density of solid carbon, approximately 0.17 moles per cubic centimeter (2.1 grams per  $\text{cm}^3$ ). Note that this is different from the pseudo-gas concentration (or density)  $\bar{\rho}_c$ , which is the amount of carbon found in one cubic centimeter of the particle laden gas. The total carbon surface area per unit volume,  $\bar{\sigma}_c$ , is given by:

$$\bar{\sigma}_c = \pi \eta d^2 \quad (4)$$

Equations 3 and 4 can be used to find  $\bar{\sigma}_c$ , the carbon surface area per unit volume, in terms of the pseudo-gas concentration of carbon,  $\bar{\rho}_c$  and the particle diameter:

$$\bar{\sigma}_c = \frac{6}{\rho_c d} \bar{\rho}_c \quad (5)$$

If a reasonable estimate can be made for the diameter of the solid carbon particles, then Equation 5 can be used to convert the pseudo-gas carbon concentration into the reacting surface area per unit volume. Next, suppose that the reaction rate for solid carbon is given in the form of Equation 2, with carbon in the place of the previously hypothetical solid reactant B. Let the gaseous reactant continue to be denoted by A, with the understanding that A will eventually stand for either  $\text{CO}_2$ ,  $\text{OH}$ , or  $\text{O}$ . Then, the reaction rate is

$$k' = \bar{\rho}_A \left( \frac{6\bar{\rho}_C}{\rho_C d} \right) \quad (6)$$

A slight rearrangement of Equation 6 yields:

$$\left( \frac{6k'}{\rho_C d} \right) = \bar{\rho}_A = \bar{\rho}_C \quad (7)$$

Equation 7 is identical to Equation 1, provided that the effective reaction rate for the carbon pseudo-gas is set equal to

$$k = \left( \frac{6k'}{\rho_C d} \right) \quad (8)$$

This reaction is in a form that can be used by SPF.

A degree of approximation is involved. In reality, the particle diameter will decrease as the particles are consumed, and thus the particle diameter will not be constant. However, the rates computed by this method should at least be the right order of magnitude.

#### Log-Normal Particle Size Distribution

Suppose the carbon particle diameters are not all the same, but are statistically distributed. The particle size distribution function is defined in the same way as a probability distribution function: the fraction of the total number of particles with diameters between  $d_1$  and  $d_2$  is equal to the integral of  $\phi(d)$  from  $d_1$  to  $d_2$ . There are several standard distribution functions which are commonly used to describe particles. One of the most common is the log-normal. This distribution function can be written in the form:

$$\phi(d) = \frac{\delta d_m^3}{\sqrt{\pi} d^4} \exp(-9/4\delta^2) \exp(-\delta^2 y^2) \quad (9)$$

where

$$y = \ln(d/d_m) \quad (10)$$

Here  $d_m$  and  $\delta$  are parameters that describe the distribution function, corresponding to a characteristic diameter and a standard deviation, respectively.

Assuming a log-normal particle size distribution, equations can be derived, corresponding to the ones in the previous section, to yield an effective reaction rate. The analogue of Equation 3 is:

$$\bar{\rho}_c = \frac{\pi}{6} \rho_c \eta d_m^3 \exp(-9/4\delta^2) \quad (11)$$

and, in place of Equation 4:

$$\bar{\sigma}_c = \pi \eta d_m^2 \exp(-2\delta^2) \quad (12)$$

Then, corresponding to Equation 8,

$$k = \left( \frac{6k'}{\rho_c d_m} \right) \exp(1/4\delta^2) \quad (13)$$

Equation 13, which was derived assuming a log-normal distribution, is nearly the same as Equation 8, which was derived assuming that all particles are the same diameter. The only differences are that the "mean" diameter  $d_m$  appears in place of the particle diameter, and the appearance of the additional multiplying factor  $\exp(1/4\delta^2)$ .

### Results of Calculations

Before evaluating the effects of cart on particle combustion, a few additional required reactions were added. SMIRF uses SPF to perform the reacting plume calculation, and SPF gets its reaction rates from an input file called DATBANK. Although the entire DATBANK file will not be reproduced, Table 3 contains an excerpt from the DATBANK file which shows the reactions of interest.

When these calculations were first begun, the magnesium vapor in the flare plume was not burning as it mixed with the oxygen in the air. Thus, reaction number 168 was added. It was also necessary to add the condensation reactions 172 and 173. The rates assigned to these reactions are not measured reaction rates; the numbers were determined

by running the flare model again and again with faster and faster rates for these reactions until the reactions seemed to be in equilibrium. Thus, the results presented here are based on the assumption of "fast" chemistry (i.e., equilibrium) for reactions 168, 172, and 173.

Table 3. Excerpt from DATBANK File								
N	Reaction				T	A	n	E
166	C	+O2	=CO	+O	61	1.0E-25	0.0	-5000.
167	MGF	+MGF	=MGF2	+MG	18	3.5E-25		-5180.
168	MG	+O2	=MGOS	+O	15	1.0E-09	0.0	-16500.
169	CGR	+O	=CO		48	3.3E-16	-0.5	0.
170	CGR	+OH	=CO	+H	18	1.8E-16	-0.5	0.
171	CGR	+CO2	=CO	+CO	18	1.3E-15	-1.0	-41926.
172	MGO	+MGO	=MGOS	+MGOS	11	1.0E-08		
173	MGF2	+MGF2	=MGFL	+MGFL	11	1.0E-10		
<p>The numbers in column N are the reaction numbers. T is a two digit code which specifies the reaction type. The documentation for SPF should be consulted for more details.<sup>5</sup> The numbers A, n, and E describe the variation of the rate constant with temperature according to the Arrhenius equation: <math>k = A T^n \exp(E/RT)</math>. T is the temperature and R is the ideal gas constant. The activation energy E is in kcal/mole; n is dimensionless. The units of A vary according to the order of the reaction, but may be found for any given reaction from the specification that the concentrations are in molecules per cubic centimeter and the resulting reaction rate is in molecules per cubic centimeter per second.</p>								

Next, the reactions for the carbon particles were added. These reactions are numbers 169, 170, and 171 in Table 3, and express the reaction of the solid carbon with the three species which were previously found to be most important. The rates for these reactions were calculated using the surface reaction rates from Table 2 for O, OH, and CO<sub>2</sub> together with Equation 13. The diameter  $d_m$  was assumed to be one micron, and the distribution parameter  $\delta$  was 1.1. The reaction rates input to SPF via DATBANK are in the molecule-cm-sec system of units, so an appropriate conversion factor was applied. The resulting reaction rate constants are given in Table 3 for reactions 169, 170, and 171.



There are no reliable particle size measurements for these flares. However, one micron seems to be a reasonable estimate for the diameter of the carbon particles, and thus the reaction rates in Table 3 for the solid carbon should at least be of the right order of magnitude. Furthermore, when the flare model is applied to actual flares for which signature measurements are available, this assumed particle diameter might be varied to improve the fit.

Figure 1 illustrates the variation of concentration as a function of distance from the centerline of the plume ( $Y$ ). This distance was made dimensionless by dividing by  $R$  (here equal to 0.13 feet). The radial profile of Figure 1 is at a station 5 feet aft of the flare. The calculation was made for a flare which has just been deployed. This result was computed using the carbon particle reaction rates from Table 3, assuming a one micron diameter.

Figure 2 shows the temperature profile in the plume. Here, the temperature profile is calculated for three different scenarios. First, the code was run with the carbon particle reactions omitted. Next, the particles were allowed to burn, using the reaction rates in Table 3, which assume a one micron particle diameter. The purpose was to examine the effect produced by allowing the carbon particles to burn. Then, the code was run a third time, only with the rates for reactions 169, 170, and 171 increased by 5 orders of magnitude. These greatly increased reaction rates were intended to simulate "fast" reactions and produce an equilibrium condition. Thus, these calculations were intended to explore the range of possibilities from no particle reactions at all to fast reactions.

The temperature profile in Figure 2 shows some signs of numerical instability (e.g., the kink in the curve for no particle reactions at  $Y/R$  of 0.5). The calculations were repeated (results not shown) with a smaller step size, and a change was noted in the region from  $Y/R$  of 0.4 to 0.7. Thus, the variation in the temperature profiles in this region may not be significant. This area requires more work. However, the temperature profiles in the rest of the plume do appear to be correct. The solid carbon particle reactions appear to make little difference in the maximum temperature. The only real effect of these reactions seems to be an increase in the width of the profile.

Figure 3 shows the variation of the mass fraction of solid carbon with radius. With no particle reactions, the decrease in solid carbon density toward the outer edge of the plume is due entirely to dilution. When the carbon reactions are enabled, the effect is clearly seen in the form of a decrease in the carbon concentration. This effect occurs again when the reaction rates are increased.

Figure 4 shows the concentration of carbon monoxide. With no carbon particle reactions, there is no pathway for the formation of carbon monoxide, so the concentration is exactly zero everywhere in the plume. When the reactions are enabled, carbon monoxide is formed where the carbon particle laden flare products mix with the air.

Figure 5 shows a similar effect on the concentration of carbon dioxide. The peak carbon dioxide concentration occurs at a larger radius than the peak carbon monoxide concentration. This can be interpreted as the result of a two step oxidation process. Under very fuel rich conditions, the monoxide is formed, which then reacts further when it reaches the outer portion of the plume where oxygen is more plentiful. Since most of the heat released by burning carbon is due to the second oxidation, this would explain the changes seen in the temperature profile (Figure 2).

The most important result is given in Figure 6, which shows the plume signature computed for these three different reaction scenarios. The radiation intensity was spectrally integrated from 2000 to 5000  $\text{cm}^{-1}$  and spatially integrated over the plume. This calculation was repeated at one second intervals from ignition to burnout, resulting in a plot of the total emissive power versus time.

Figure 6 demonstrates the effect of including the carbon particle reactions. Going from no carbon particle reactions at all to the fast reaction (equilibrium) limit makes an order of magnitude difference in the integrated total emissive power of the plume. Using the computed reaction rates for one micron carbon particles, the answer falls between these two extreme limits.

## Conclusions

A method has been found which allows the inclusion of reacting particle effects in existing computer programs which are written to handle chemically reacting gases. This approximation treats the particulate species as pseudo-gases, and computes effective rates for reactions involving these species based on an assumed particle diameter.

The plumes of Teflon-magnesium flares contain large quantities of unreacted carbon in the form of solid particles. Reactions between these solid carbon particles and the gaseous species O, OH, and CO<sub>2</sub> were added to the SMIRF flare model. The results indicate that these gas-particle reactions may make an order of magnitude difference in the predicted flare signature.

The flare signature calculated by using the actual reaction rates for the particulate reactions is clearly different from the signature calculated by either ignoring these reactions, or by making them so fast as to be in equilibrium. Thus, accurate prediction of flare signatures will require that the kinetics of these reactions be modelled. They cannot be ignored, and they cannot be assumed to be fast.

## Acknowledgements

This research was performed at Arnold Engineering Development Center (AEDC) under the 1991 Air Force Office of Scientific Research (AFOSR) Summer Faculty Research Program, administered by Research and Development Laboratories (RDL) of Culver City, California. I wish to thank the Air Force Systems Command and AFOSR for their sponsorship. RDL must be acknowledged for their capable administration of this program, as should the coordinator of the summer program at AEDC, Major David Hart. This summer research effort was guided by H. T. Bentley III of Overdrup Technology. Acknowledgement is also due to J. C. Denny and Martha Simmons of Overdrup Technology, and Dr. Robert P. Rhodes of the University of Tennessee Space Institute.

## References

1. Denny, J. C. and D. G. Brown, *SMIRF, Signature Model for Infrared Flares, with Post Processor*, April 19, 1991.
2. Gordon, S. and B. J. McBride, *Computer Program for Calculation of Complex Chemical Equilibrium Compositions, Rocket Performance, Incident and Reflected Shocks, and Chapman-Jouguet Detonations*, NASA/Lewis Research Center, NASA SP-273, 1976.
3. Norton, O. P., *Combustion of Carbon Particles in the Plume of a Flare*, final report for 1990 AFOSR-UES Summer Faculty Research program, AFOSR contract number F49620-88-C-0053, 1990.
4. Norton, O. P. and H. T. Bentley, III, *Combustion of Carbon Particles in the Plume of a Flare*, presented at the 1991 Meeting of the IRIS Specialty Group on Infrared Countermeasures, held May 14-16, 1991 at the Johns Hopkins Applied Physics Laboratory.
5. Dash, S. M., H. S. Pergament, D. E. Wolf, N. Sinha, M. W. Taylor, and M. E. Vaughn, Jr., *The JANNAF Standardized Plume Flowfield Code Version II, (SPF-II)*, volumes I and II, Technical Report CR-RD-SS-90-4, U. S. Army Missile Command, July, 1990.

Plume Concentration Profile  
Generic Flare, T=0 sec, X=5 ft

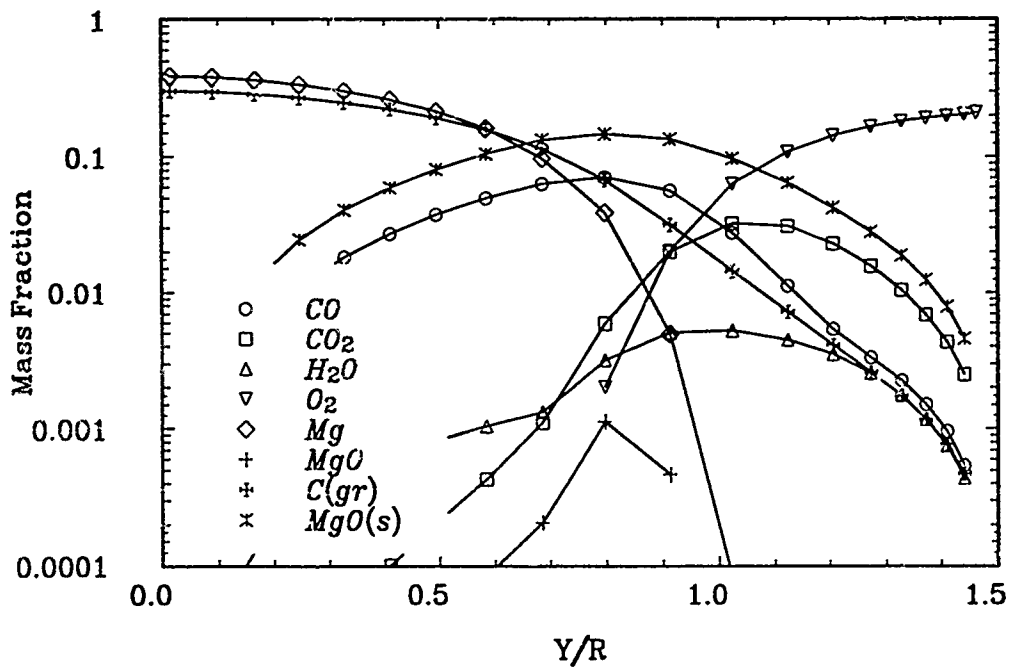


Figure 1. A plot of the concentrations of selected species as a function of radial distance from the centerline of the plume. The flare has just been released and ignited ( $t=0$  sec), and this cross section of the plume is at a distance 5 feet downstream from the flare ( $x=5$  ft). The species  $C(gr)$  is graphite, or solid carbon.

Plume Temperature Profile  
Generic Flare,  $T=0$  sec,  $X=5$  ft

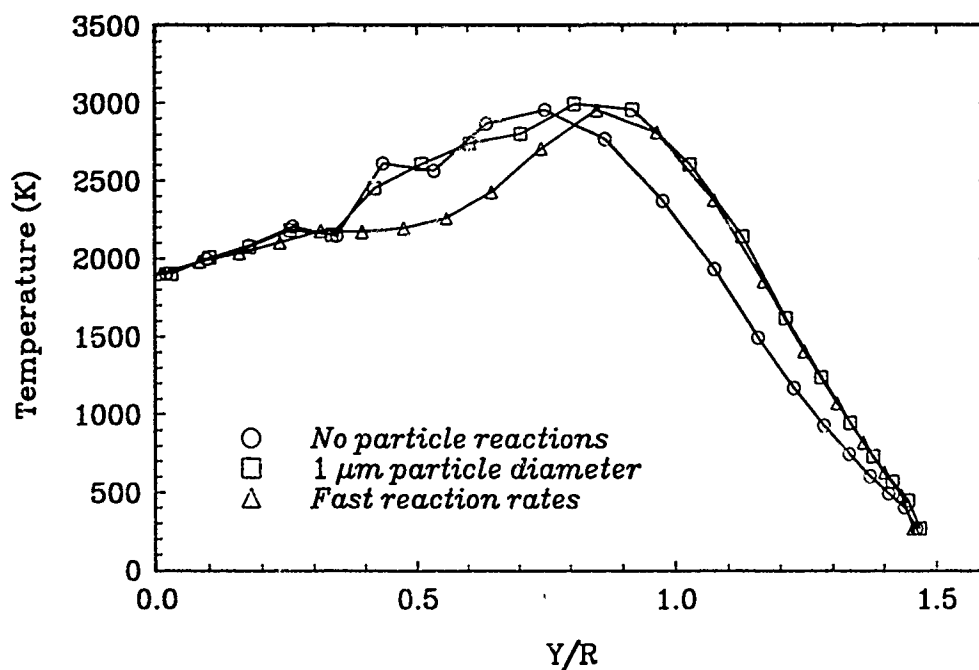


Figure 2. Under the same conditions as in Figure 1 ( $t=0$  sec,  $x=5$  ft), the radial temperature profile is shown. Three different curves are shown corresponding to no reactions for carbon particles, reaction rates calculated based on a 1 micron particle diameter, and extremely fast reactions (equilibrium).

$C(\text{gr})$  Concentration Profile  
Generic Flare,  $T=0$  sec,  $X=5$  ft

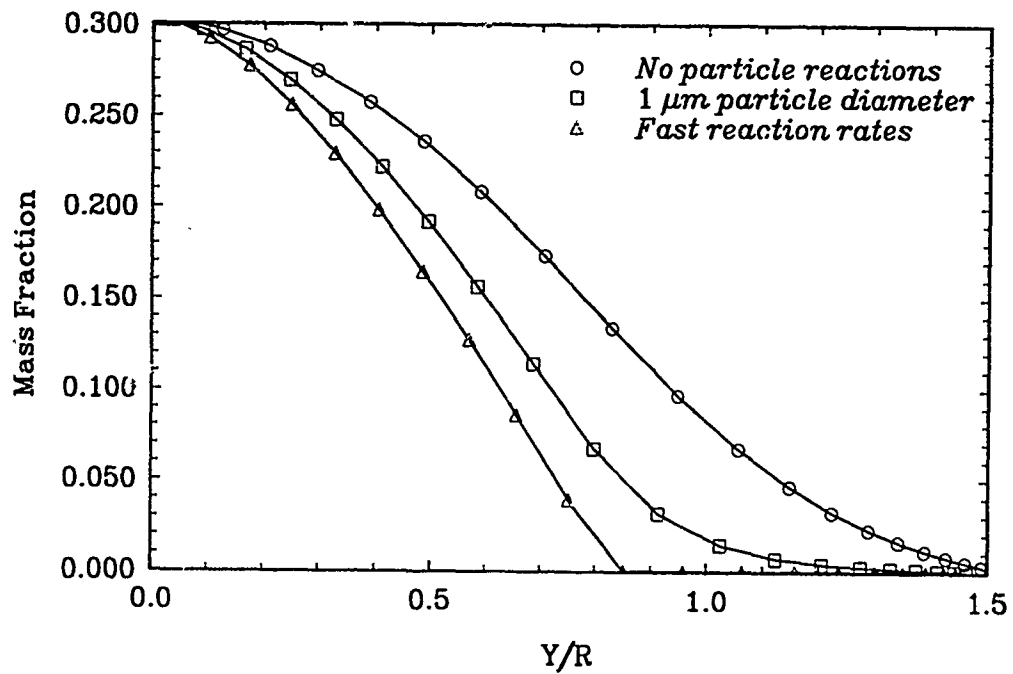


Figure 3. A plot of the carbon particle density in the plume. As before, this plot corresponds to  $t=0$  seconds and  $x=5$  feet.

CO Concentration Profile  
Generic Flare, T=0 sec, X=5 ft

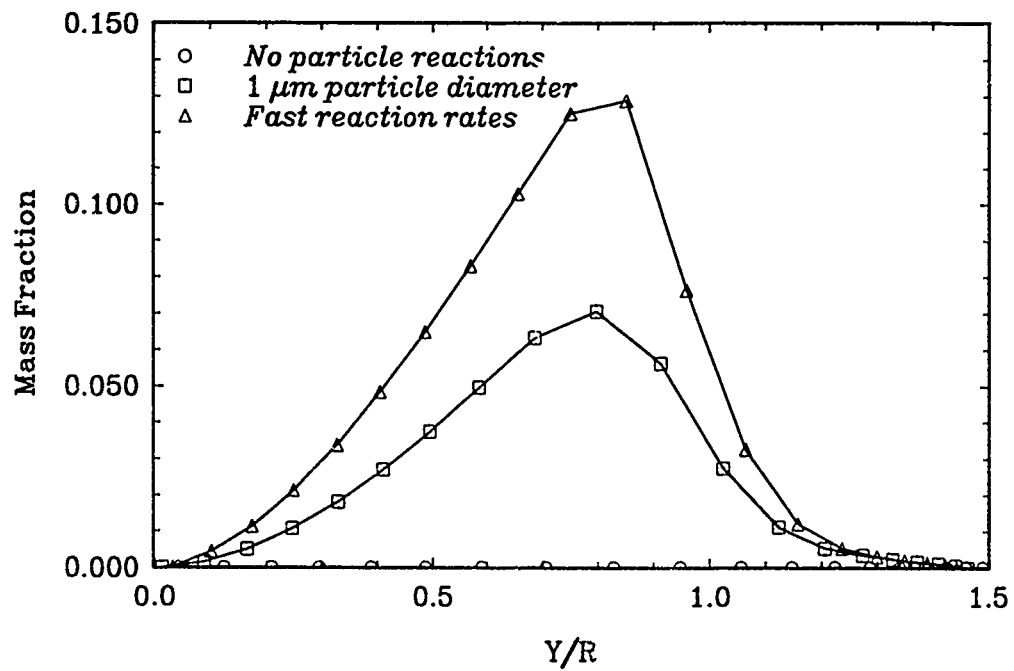


Figure 4. A plot of the concentration of carbon monoxide in the plume. This plot corresponds to  $t=0$  seconds and  $x=5$  feet.



CO<sub>2</sub> Concentration Profile  
Generic Flare, T=0 sec, X=5 ft

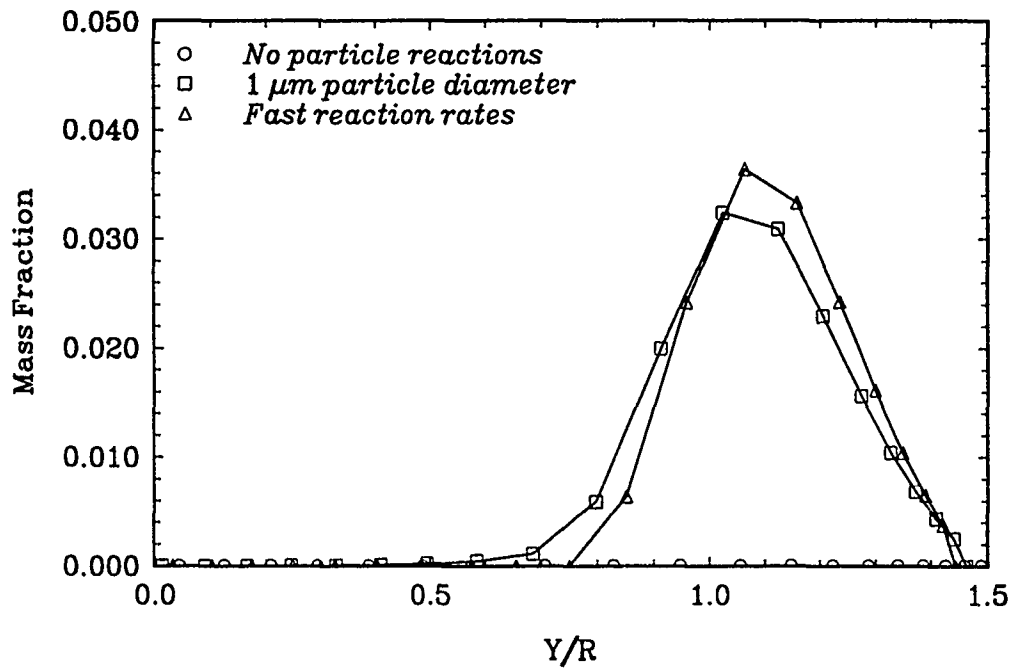


Figure 5. A plot of the concentration carbon dioxide in the plume. This plot corresponds to  $t=0$  seconds and  $x=5$  feet.

Flare Model Post Processor Output  
Generic Flare

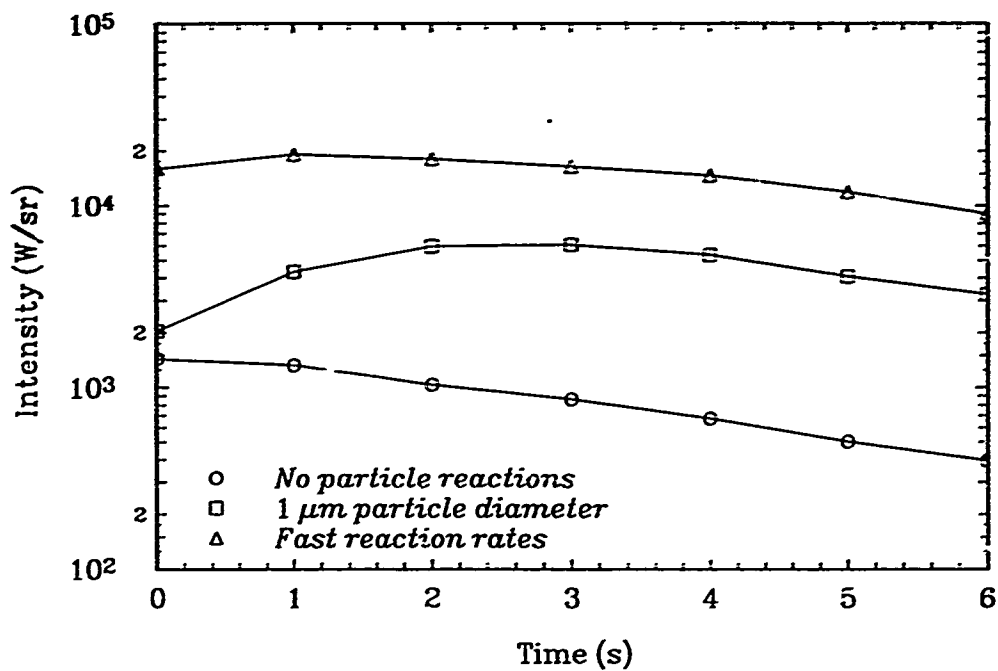


Figure 6. A plot of the intensity of infrared emissions from the flare as a function of time. The emissions have been integrated spatially and spectrally (from 2000 to 5000  $\text{cm}^{-1}$ ).

**1991 USAF-RDL SUMMER FACULTY RESEARCH PROGRAM  
GRADUATE STUDENT RESEARCH PROGRAM**

**Sponsored by the  
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH  
Conducted by the  
Research and Development Laboratories  
FINAL REPORT  
Software for 2D and 3D Mathematical Morphology**

<b>Prepared by:</b>	<b>Richard Alan Peters II, Ph. D.</b>
<b>Academic Rank:</b>	<b>Assistant Professor</b>
<b>Department and</b>	<b>Department of Electrical Engineering</b>
<b>University:</b>	<b>Vanderbilt University</b>
<b>Research Location:</b>	<b>Arnold Engineering and Development Center Arnold AFB, TN 37389</b>
<b>USAF Researcher:</b>	<b>J.A. Nichols</b>

## Software for 2D and 3D Mathematical Morphology

by

Richard Alan Peters II

### ABSTRACT

Mathematical morphology is a powerful tool for image analysis and enhancement. Morphological operators are shape-dependent, nonlinear image transforms such as erosion, dilation, opening, closing, and rank filters. The mathematical operators are defined in  $n$  dimensions so it is possible to create programs that will operate on 1D signals, 2D images, or 3D datasets using exactly the same concepts. As operators on 2D binary images, mathematical morphology is well known; much software is available which performs binary morphology. However, the true power in morphology lies in its ability to transform grayscale 2D pixel images and 3D voxel datasets. Yet there is no commercially available software that performs these functions.

This is a description of a morphological software package written under the auspices of the US Airforce Office of Scientific Research at Arnold Airforce Base during the 1991 Summer Faculty Research program. The software includes a 2D image morphology program, a 3D voxel image morphology program, a program for enhancement and noise reduction of 2D images, and related support routines for image arithmetic and logical operations. These programs will perform many - if not all - of the possible morphological operations on both binary and grayscale images. They provide many of the commonly used features automatically or permit the user to customize the operation to fit an application. The user programs operate on Sun rasterfiles. However, the main morphological routines are c-code subroutines. These are independent of specific file formats, so a users can easily write an interface program to operate on other file formats.

### Acknowledgements

I would like to acknowledge the Air Force Office of Scientific Research and the Research and Development Laboratories for conducting the Summer Faculty Research Program. It has been an excellent opportunity for me to work unhindered on tools that will prove useful for research in the years to come. Thanks are due to the staff of Arnold Engineering and Development Center, Including Maj. Hart, for making my tenure free of bureaucratic responsibility so that I could devote my full time and effort to the engineering problems at hand. I offer my sincerest thanks to Jim Nichols for his excellent suggestions and guidance. AEDC is extraordinarily fortunate to have him working for them.

## I. INTRODUCTION

Mathematical morphology is the name applied to a collection of set theoretic operators defined on a abstract structure known as an infinite lattice. These operators were first examined systematically by Matheron [6] and Serra [8], [9], [10] in the 1960's, are an extension of Minkowsky's set theory. They are especially useful for image analysis and image enhancement. There are, in the general literature, a number of good tutorials on image morphology [1], [2], [3], [4], [5], [9]. We refer the novice to any of these papers for the fundamentals.

Morphological operators include, erosion, dilation, opening, closing, rank filters (includes median filter), tophat transforms, and generalized correlations. These operations can be defined on binary or grayscale images in any number of dimensions. A morphological operator is governed by a small pseudo image called a "structuring element" (SE). When applied to an image, the operator returns a quantitative measure of the image's geometrical structure in terms of the SE. This measure can be used to isolate features in an image or to construct a synthetic image containing regular approximations of the features in the original. A "regular" feature, in this sense, is a feature that can be reconstructed uniquely by the operator using the SE.

Certain morphological operators, when applied to 2D grayscale images or 3D voxel images return a synthetic image that is - in a sense - a smoothed version of the original image. But, the smoothing is nonlinear. Unlike a linearly smooth image, a morphologically smooth image usually contains sharp edges. The pieces of the original image that are not present in the transformed image are those features that could not be approximated by the SE. These pieces are in the residual, the difference between the original image and the smoothed version. Both the smooth output and the residual can be quite useful. The smooth output is, in a sense, a root or core image. It is simpler than the original while retaining important characteristics. This can make analysis simpler. The residual contains all "irregular" features such as noise and edges that could not be approximated by the SE. With a few assumptions about the differences between noise and edges, one can remove much of the noise from the residual image. Adding the "cleaned-up" residual back to the regular features yields a relatively noise free image with many edges intact and unblurred. (Note: We have included a program, *mclean*, in our package, which does precisely that.)

Mathematical morphology extends to the analysis of 3D data sets in two ways. Morphological operators applied to three dimensional voxel images such as the output from a CT, MRI or PET scanner are completely analogous in effect to those applied to 2D imagery. A morphological operator will approximate within specific 3D shape constraints, the density or intensity of voxels in space. Thus, nonlinear smoothings and residuals can be constructed for analysis or enhancement of a 3D image.

Three dimensional morphological operators are quite useful in another context. Consider that a time sequence of 2D images can be thought of a 3D image in spacetime (2 space axes, 1 time axis). If we treat time as a spatial dimension, we can apply 3D morphology to a time sequence of images. The properties of spacetime can be exploited by morphological processing for the analysis and enhancement of moving imagery.

A true object (not an artifact of imaging) will exist in multiple frames of a time sequence, as long as it is not moving too rapidly through the field of view. Noise, however, and other spurious artifacts of the imaging process will change, often dramatically, from frame to frame. An appropriate 3D morphological operator, applied to such a sequence can preserve the persistent shapes and discard those which change rapidly. This researcher has observed that simple 3D morphological operations applied to a noisy time sequence can reduce dramatically the apparent noise.

Two dimensional objects in a time sequence of images become 3D generalized cylinders, or

"worms" in 3D spacetime. The pitch or angle these worms make with respect to the time axis is directly related to the velocity with which the objects are moving in the time sequence. It is a simple matter to design structuring elements that are sensitive to specific velocities or ranges of velocities since it is simple to specify the shape of the SE. Thus, it is possible to use 3D morphology to enhance or filter out objects moving with specific velocities in an image sequence.

Mathematical morphology is of unquestionable utility in the analysis and enhancement of very noisy imagery like that processed by scientists and engineers at Arnold Engineering and Development Center. This software will give them the capacity to perform both 2D and 3D morphology on their imagery.

## II. OBJECTIVES OF THE RESEARCH EFFORT

The objective of this research effort was to provide the scientists and engineers at Arnold Engineering and Development Center with robust, reliable, and easy to use software for two and three dimensional mathematical morphology. The programs had to be

- **Comprehensive.** It should implement virtually every commonly used morphological operation. It should operate on both binary and grayscale imagery. It should implement both set and function morphology and hit-or-miss transforms. It should be able to use structuring elements of any size, shape, or connectivity.
- **Flexible.** It should be easily extensible to add new functionalities or new structuring elements. It should be easily modifiable to operate on images in a variety of formats. It should run on any Unix workstation with minimal changes.
- **Robust.** It should be as bug-free as possible. It should be predictable; similar inputs under similar operators should produce similar outputs. Novel uses within reason should not crash the program.

The author feels confident that he has met the first objective. While it is trivial to add new structuring elements, the second objective is not as straightforward to accomplish in every respect. It was not possible to write a specific user interface for the program and have it read any image file format. The author got as close as possible to that ideal by writing the main 2D and 3D morphology procedures as subroutines, decoupled from any specific file type. He also wrote user interface - I/O programs that read and write Sun rasterfiles and pass data to the subroutines. Users wishing to use other file formats will either need to use file conversion programs, like pbmplus or image alchemy, to convert their formats to rasterfiles or they will need to write their own file I/O program that calls the morphological subroutines. The programs have been written making no use of SunOS-specific system calls. They do not employ a graphical user interface - they use a simple command line interface. This should ensure portability into most unix environments. In fact, the programs should run with few changes on any PC that has linearly addressable memory such as an Amiga or Macintosh, providing it has sufficient memory.

### III. 2D IMAGE MORPHOLOGY SOFTWARE

In this section we will detail the use of *MorphSub*, the 2D morphology subroutine, and *morph*, its Sun rasterfile implementation.

#### MorphSub

Subroutine *MorphSub* performs many (if not all) possible 2D morphological operations on gray-level or binary images. This includes hit-or-miss transforms, order statistic filters, function processing or set processing erosion, dilation, opening and closing, tophat (image minus opening), bothat (closing minus image), and isolated delete functions.

To use make the following call in a c program:

```
MorphSub( MOp, In, Out, Ix, Iy, ImgType, NZPad, LThresh, UThresh,
          SENAME, AutoSE, Sx, Sy, Sz, SEType, SorF, Rank, NoScale, Dsp)
```

where the parameters are defined by:

int	MOp;	Morphological operation to perform
byte	*In;	input image as byte list in row major order
byte	*Out;	output image as byte list in row major order
int	Ix, Iy;	image horizontal, vertical dimensions
int	ImgType;	image type (gray-level or binary)
int	NZPad ;	flag. F $\Rightarrow$ zeropadding of input
int	LThresh;	lower binary threshold value
int	UThresh;	upper binary threshold value
char	*SEName;	structuring element (SE) path name
int	AutoSE;	use a canned or auto generated SE
int	Sx,Sy,Sz;	SE x and y support dims and max gray-lev (z)
int	SEType;	Binary SE or gray-level SE
int	SorF;	Set operation or Function operation
int	Rank;	rank for rank filter
int	NoScale;	flag. T $\Rightarrow$ do not scale output of IntMorph
int	Dsp;	flag. T $\Rightarrow$ display some info

A detailed explanation of parameters follows:

Mop Morphological Operation. This integer parameter can have the following values:

mnemon	hex	dec	action
ERODE	0x0001	1	erosion
DILATE	0x0002	2	dilation
OPEN	0x0004	4	opening
CLOSE	0x0008	8	closing
RANK	0x0010	16	rank filter (order statistic filter)
TOPHAT	0x0020	32	tophat (image minus opening)

<b>BOTHAT</b>	<b>0x0040</b>	<b>64</b>	<b>bothat (closing minus image)</b>
---------------	---------------	-----------	-------------------------------------

Moreover, the erode and dilate functions can be "negated" using:

<b>mnemon</b>	<b>hex</b>	<b>dec</b>	<b>action</b>
<b>NOTFLG</b>	<b>0x0080</b>	<b>128</b>	<b>"not" flag for binary erodes and dilates</b>

Let I = original image; E = eroded image; D = dilated image. The following are valid parameters:

<b>mnemon</b>	<b>hex</b>	<b>dec</b>	<b>action</b>
<b>ERODE   NOTFLG</b>	<b>0x0081</b>	<b>129</b>	<b>I &amp;&amp; !E (pixelwise)</b>
<b>DILATE   NOTFLG</b>	<b>0x0082</b>	<b>130</b>	<b>D &amp;&amp; !I (pixelwise)</b>

**ERODE | NOTFLG** with a binary hit-or-miss structuring element will delete in a binary image, white features with the shape of the "hit" portion of the SE. (e.g. one can easily devise a SE to delete isolated white pixels). With a gray-level SE it will delete the "interiors" from sets of white pixels in a binary image.

**DILATE | NOTFLG** with a binary hit-or-miss structuring element will delete in a binary image, black features with the shape of the "hit" portion of the SE. (e.g. one can easily devise a SE to delete isolated black pixels). With a gray-level SE it will delete the "interiors" from sets of black pixels in a binary image.

- In**     The input image. A pointer to an Ix-times-Iy long list of bytes. The byte list is a one-byte per pixel grayscale image that is Iy rows by Ix columns in row major order. (First row of pixels, followed by second row, followed by third, ...).
- Out**   The output image. A pointer to an Ix-times-Iy long list of bytes. The byte list is a one-byte per pixel grayscale image that is Iy rows by Ix columns in row major order. This is where morph will write its output. The space must be allocated by the calling program.

Out and In may point to the same space without error. However, morph will then overwrite the input image with the output.

**Ix**     Image horizontal dimension. Number of pixels per line (or number of columns per row) in image.

**Iy**     Image vertical dimension. Number of lines (or rows) in the image.

**ImgType**   Image type: This specifies whether the input image is to be treated like a gray level or binary image. The values in a binary image are zero and not-zero (it does not matter which of the values in [BLACK,WHITE]).

Parameter values:



mnemon	hex	dec	meaning
GRAIMG	0x0000	0	gray-level image
BINIMG	00x0200	512	binary image

**NZPad** If zero (FALSE) *MorphSub* will extend the size of the input image internally and create a border of zeros around it. If nonzero (TRUE) *MorphSub* will not do the zero padding. Note: this is all internal to *MorphSub*. The input and output images are both *Ix* by *Iy* independent of this flag.

There are image border effects with each option. The effects differ depending on the option. With the option FALSE, *morph-sub* does what it can to "color in" the image near its perimeter. With this option TRUE, the output image has a border of zeros inside it the width and height of the SE. That is, the transformed area of the output image is smaller than the actual image dimensions. This is, in a sense, a more accurate result than the zero padded default. To zero pad the input permits the program to transform the border region, but it does this on the assumption that the original scene was black outside the image. This, of course, is almost never true. Thus, the border region is inaccurately transformed. Use this switch if accuracy is more important than having an image that is "colored in" out to the boundary.

**LThresh** lower binary threshold.

**UThresh** upper binary threshold.

These two parameters will extract a binary image from a grayscale image via thresholding if: UThresh is non zero AND SorF is SET. Then all pixels with grey-levels in [Lthresh,UThresh] will be set to WHITE, all others to BLACK. Thresholding occurs before any morphology. This makes it possible to specify binary morphology on a graylevel image. If UThresh > 0 AND LThresh > Uthresh, *MorphSub* aborts with an error. If UThresh is zero no thresholding is performed. Note that one can use *MorphSub* to perform a simple threshold on an image by specifying MOp = ERODE, ImgType = GRAIMG, AutoSE = AUTO Sx = 1, Sy = 1, Sz = 0, and the appropriate LThresh and UThresh.

**SEName** This is a string pointer to the file (or path) name of a structuring element (SE) file. If this pointer is not NULL and AutoSE==0, then SEName is catenated to the end of the value of environment variable SEPATH to get the SE file pathname. (If SEPATH is not defined, then SEName is used alone.) See below for more complete info on SE's.

**AutoSE** If this flag is non zero, it signifies: use a canned or auto generated SE. Its possible values are:

mnemon	dec	meaning
ZERO	0	do not use auto SE
AUTO	1	generate disk shaped SE
PLUS	2	use 3 by 3 "+" shaped SE
S3X3	3	use 3 by 3 square SE
S5X5	4	use 5 by 5 quasi disk SE

See below for more complete info on SE's.

- Sx,Sy** Horizontal and vertical dimensions of program generated SE. Ignored if AutoSE != AUTO.
- Sz** Gray level of origin pixel in program generated SE. This is used if AutoSE == AUTO. Ignored if AutoSE != AUTO.
- SEType** specifies whether the structuring element is to be interpreted as a gray-level SE or a binary SE. (See below for the distinction.) Its possible values are:

mnemon	hex	dec	meaning
GRASE	0x0000	0	gray-level SE
BINSE	0x0100	256	binary SE

- SorF** Specifies whether the morphological operation is to be of the "set" type or "function" type. (See below for the distinction.) Its possible values are:

mnemon	hex	dec	meaning
SET	0x0000	0	set operation
FUNCT	0x0400	1024	function operation

#### Structuring element specifications:

**SEFile:** A structuring element file is an ASCII file of integers separated by spaces. The first two numbers,  $x$ ,  $y$ , are the horizontal and vertical dimensions in pixels of the smallest rectangle that will cover the structuring element. Both  $x$  and  $y$  must be  $> 0$ . The next two numbers,  $i$ ,  $j$  are the horizontal and vertical coordinates, respectively, of the SE origin. **IMPORTANT:** The origin is expected to be in the covering rectangle. If not, *MorphSub* aborts. The upper left hand corner of the rectangle has coordinates  $(0, 0)$ ; the lower right is  $(x-1, y-1)$ . Following the first four integers are  $x \times y$  integers separated by spaces. These numbers are the SE elements. Their interpretation depends on the morphological operation being performed.

Negative SE elements are ALWAYS treated as logical DON'T CAREs. That is, when the operation is in progress, image pixels under negative SE elements are ignored. Thus, the support of the SE is limited to those elements that are nonnegative. This permits the creation of odd-shaped and multiply connected SE's or the placement of the SE origin outside the body of the SE. If *ImgType* is binary, (i.e. pixels grouped as zero and not zero), and if *SEType* is binary, then the SE is used to perform a hit-or-miss transform. In this case, zero SE elements cover the "miss" support and positive (nonzero) elements cover the "hit" support. The actual gray-levels are ignored.

If *ImgType* is binary, and *SEType* is gray then the nonnegative (both zero and greater than zero) SE elements determine the support of a "hit-only" transform. That is, the nonnegative support is used as a standard set-type SE for set (binary) morphology. (Of course, the other gray-level info is ignored.) Note: if *ImgType* is binary, then a set operation is performed by default (*SorF* is ignored).

The interpretation of the SE elements for `ImgType` gray depends on the flags `SEType` and `SorF`:

<code>SEType</code>	<code>SorF</code>	action
binary	set	Function-set morphology on support of strictly greater than zero SE elements.
binary	funct	Same as above.
gray	set	Function-set morphology on support of nonnegative (greater than or equal to zero) SE elements.
gray	funct	Function-function morphology on support of nonnegative SE elements.

**AUTO:** the program makes an SE. The self-made SE is a disk with a diameter of `Sx` pixels horizontally and `Sy` pixels vertically. `Sx` and `Sy` must be odd and greater than or equal to 1. If `AutoSE == AUTO`, `SEType` is set to gray (`SEType == GRASE`). `Sz` is the gray level of the center pixel. If `Sz > 0`: the SE will have a "curved" top. (Use `Sz > 0` for a rolling ball transform.) A function operation (`SorF == FUNCT`) is performed. If `Sz == 0`, the SE is flat-topped. A set operation is performed (`SorF == SET`).

**S3X3:** specifies the SE to be a 3 by 3 square of pixels.

**PLUS:** specifies the SE to be a 3 by 3 "+" shaped set of pixels.

**S5X5:** specifies the SE to be a 5x5 quasi disk (square without corners) of pixels.

If one of the canned SE's (3x3, plus, or 5x5) is chosen, then `SorF` is set to SET and `SEType` is set to GRASE.

Note that when `AutoSE != 0`, the values of `SEType` and `SorF` are ignored.

**Rank** This is meaningful only if `Mop == RANK`. Then the parameter indicates the order of the filter. The rank option of *MorphSub* is actually an order statistic filter. To compute the order, *MorphSub* counts the number of pixels in the support of the SE. In a gray-level SE that is all values greater than or equal to zero. We did not define a hit-or-miss rank filter. Therefore, to compute the support of a binary SE, *MorphSub* changes the zeros in a binary SE into -1's (DON'T CARES). Then all the rank routine considers the support to be all SE elements  $\geq$  ZERO.

If `Rank == 0`, *MorphSub* applies a median filter. If `Rank == 1` it does a dilation. If `Rank == support of SE`, then *MorphSub* does an erosion. Otherwise, *MorphSub* does an order statistic filter of order `Rank`.

**NoScale** When a function processing operation is performed on a gray-level image with a gray-level SE, the operation is performed with 16-bit-precision signed arithmetic. When

the operation is through, if NoScale == 0, *MorphSub* scales the result to fit into one byte per pixel. If NoScale != 0, morph limits the results at BLACK and WHITE.

Scaling is performed as follows: if the range of the result is less than one byte (i.e.  $\text{MaxPixVal} - \text{MinPixVal} < \text{or} = 256$ ) the result is translated so that the minimum is zero. (i.e. each resultant pixel is replaced with  $\text{pixel} - \text{MinPixVal}$ ) Otherwise, the minimum is subtracted from each pixel and the difference is scaled by the ratio  $\text{WHITE}/(\text{MaxPixVal} - \text{MinPixVal})$ .

Dsp If this is true (nonzero) some info about what is going on is displayed.

## morph

The user program *morph* takes parameters from the user on the command line and applies routine *MorphSub* to Sun rasterfiles. A brief description of *morph* follows. This description borrows heavily on the description of *MorphSub* above.

usage:

```
morph < In > Out -m e|d|o|c|r|t|b|p|q [-i g|b] [-s g|b] [-o s|f] [-t nnn [mmm]]
      [-r med|nnn] [-z] [-n] [-v] -k SEFile | 3x3 | 5x5 | plus | auto x y [z]
```

where:

In is the path name of the Sun rasterfile input image. (Type ;In so the file is read in through stdin.)

Out is the pathname of the Sun rasterfile output image. (Type ;Out so the file is output through stdout.)

*morph* extracts and operates on the luminance component of the image. Thus the output of *morph* is grayscale even if the input is color.

-m The letter following -m indicates the morphological operation:

- e - erode
- d - dilate
- o - open
- c - close
- r - rank filter
- t - top hat transform (image minus opening)
- b - bot hat transform (closing minus image)
- p - I && !E (pixelwise) where I = original image; E = eroded image;
- q - D && !I (pixelwise) D = dilated image;

one of these letters must be specified; there is no default.

p with a binary hit-or-miss structuring element will delete in a binary image, white features with the shape of the "hit" portion of the SE. (e.g. one can easily devise a SE to delete isolated pixels).

p with a gray-level SE will delete the "interiors" from sets of white pixels in a binary image.

q with a binary hit-or-miss structuring element will delete in a binary image, black features with the shape of the "hit" portion of the SE. (e.g. one can easily devise a SE to delete isolated pixels).

q with a gray-level SE will delete the "interiors" from sets of black pixels in a binary image.

- i Switch -i indicates that the next letter tells the image type: either g for a gray-level image or b for a binary image. If -i is not included, the default is gray-level.
- s Switch -s indicates that the next letter tells the structuring element type: either b for a binary SE or g for a gray-level SE. If -s is not included, the default is binary.
- o The letter following -o, either s or f, indicates that the operation is either a set operation or a function operation. (See reference.) If -o is not included, the default is set op.
- t -t nnn [mmm] indicates that a threshold of value nnn from below (and mmm from above; if unspecified mmm == 255) will be used on the input if the following 2 criteria are true: the input is a gray-level image AND the operation is a set operation. If the two criteria are true and -t nnn [mmm] is not included, the operation is treated as a function and set processing (FSP) operation (See ref.). If the criteria are not true and -t nnn is specified anyway, it is ignored. Note that you can do a simple threshold of a gray level image with:

```
morph < in.ras > out.ras -m e -t 128 -k auto 1 1
```

- r Switch -r is meaningful only if -m r is specified. Then the field following -r indicates the order of the filter. If the letters "med" are in the field, a median filter is used. If the field contains a number, nnn, then that value is used. If -op r is specified and -r med|nnn is not, the rank filter defaults to a median filter.
- z The presence of switch -z tells the program NOT to zero-pad the boundary of the image.
- n Switch -n tells the program NOT to scale the output of the operation. Such scaling happens by default for a function operation (-o f) on a gray-scale image (-i g) with a gray-scale SE (-s g). This switch is ignored by other operations.
- v Switch -v tells the program to display some info as it computes.

-k Switch -k indicates that the next field is one of 5 things:

3x3 This selects S3X3 in routine *MorphSub*.  
plus This selects PLUS in routine *MorphSub*.  
5x5 This selects S5X5 in routine *MorphSub*.  
auto  $x\ y\ [z]$  - the program makes an SE.

The self-made SE created when -k auto is specified is a disk with support covering  $x$  pixels horizontally and  $y$  pixels vertically.  $x$  and  $y$  must be odd. If specified,  $z$  is the gray level of the center pixel. If  $z \neq 0$ , a function operation (-o f) is performed. If  $z$  is not given or  $z = 0$ , the level defaults to BLACK (0) and a set operation is performed (-o s).

If the field following -k is not one of the above, then the string, called SEFILE in the usage example, is taken as the pathname of a structuring element file. If the user has an environment variable called SEPATH, the program appends SEFILE to it for the complete pathname.

If one of the canned SE's (3x3, plus, or 5x5) is chosen, the -s and -o flags are forced to be -s g -o s.

If "-k auto" is specified the SE type is forced to gray (-s g).

Included with *morph* are Unix c-shell scripts that set up *morph* to run some of the more common morphological operations. these include:

bclose	binary closing
bdilate	binary dilation
berode	binary erosion
bopen	binary opening
brank	binary rank filter
cleanup	example of image noise reduction using rolling ball
gfclose	gray-level function-function closing
gfdilate	gray-level function-function dilation
gferode	gray-level function-function erosion
gfopen	gray-level function-function opening
grank	gray-level rank filter
gsclose	gray-level function-set closing
gsdilate	gray-level function-set dilation
gserode	gray-level function-set erosion
gsopen	gray-level function-set opening
hitormiss	binary hit-or-miss transform
invroll	inverted rolling ball transform
invtop	inverted tophat transform (bothat)
medge	morphological edge detection
rollball	rolling ball transform
thresh	threshold a gray-level image
tophat	tophat transform

## **mclean**

Program *mclean* is an image enhancement / noise reduction program that uses 2-dimensional gray-scale mathematical morphology. The program cleans up any image with noise in its luminance component such that the noise has a smaller variance than the most important luminance edges.

In its default configuration, the program computes a smoothed version (S) of the luminance component (L) of the original image (I) by taking a pointwise average of the opening (O) and the closing (C) of L with a user-specified structuring element (SE). It subtracts S from L to create a residual image (R). The positive pixels in R form an image called the "tophat" (T) and the negative pixels form an image called the "bothat" (B). T contains bright image features that could not be created with the SE. (Recall that O is the best approximation of L than can be created by overlaying SE's subject to the constraint that L is pointwise everywhere GREATER than or equal to O.) T, then, contains noise as well lines, spots, and other thin or small (with respect to the SE), bright image features. B contains similar types of features, but dark rather than light. (Recall that C is the best approximation of L than can be created by overlaying SE's subject to the constraint that L is pointwise everywhere LESS than or equal to B.)

If the gray level variance of the noise in L is less than that of important luminance features such as lines and highlights, the noise can be filtered from R by inverse center clipping. That is, create a thresholded residual image (RT) by replacing with zero the value of any pixel in R whose magnitude is lower than some threshold. In general, the upper threshold, u (the threshold for T), is different in magnitude from the lower threshold, l (the threshold for B). Threshold u is nonnegative and l is nonpositive. The important features in R remain in RT along with some isolated very bright or very dark noise points (isolated single pixels). Define the support of RT to be the set of all nonzero pixels in RT. Then, the noise points are removed by performing an isolated delete on the support of RT.

RT is recombined with S as follows: Let TT be the thresholded tophat part of RT, and let BT be the thresholded bothat part of RT. Subtract u pixelwise from the TT and add the result (pixelwise) back to the S. Add l pixelwise to BT and subtract the result(pixelwise) from S. This procedure puts the important original features back in L without the noise.

Program *mclean* operates on Sun rasterfiles. It automatically extracts the luminance portion of the image if the rasterfile has a colormap. The output of *mclean* is a grayscale rasterfile. To recreate a color image requires extracting the hue and saturation components from the original image and recombining these with the cleaned up luminance. The author has not tried morphological processing on the H and S components. Also, to remake LHS image back into an 8-bit color (lookup-table indexed) image requires use of a color quantization scheme such as Heckbert's median cut. It is not known what noise this will reintroduce into the result.

### **usage:**

```
mclean InFile OutFile [-s g|b] [-o s|f] [-t|b] [-l nnn] [-u nnn] [-f nn.nn] [-r nn.nn] [-h] [-a OCFile]
      -k SEFile | 3x3 | 5x5 | plus | auto xxx yyy [zzz]
```

### **where:**

InFile    is the path name of the Sun rasterfile input image.

OutFile   is the pathname of the Sun rasterfile output image.

*mclean* extracts and operates on the luminance component of the image. Thus the output of *mclean* is grayscale even if the input is color.

- s Switch -s indicates that the next letter tells the structuring element type: either b for a binary SE or g for a gray-level SE. If -s is not included, the default is binary.
- o The letter following -o, either s or f, indicates that the operation is either a set operation or a function operation. If -o is not included, the default is set op. Selecting f along with a not-flat-topped SE will more closely approximate the original image. However, it also seems to make the noise reduction less effective.
- t The presence of -t on the command line tells *mclean* to do a tophat only process (i.e. don't use the bothat). This can be more effective than the general top/bot approach if it appears that most of the image noise and features are lighter than their surroundings. Similarly, -b means do a bothat only. This can be more effective than the combo if most of the noise and features are darker than their backgrounds.
- l Switches -l and -u cause *mclean* to use the numbers following as lower and upper thresholds, respectively. These override *mclean's* automatically determined bothat and tophat thresholds.
- f Switch -f precedes a floating point number that is used as a scale factor to adjust both the lower and upper thresholds. If it seems that *mclean* has been too extreme in its thresholding, use -f 0.nn to adjust the threshold down. If, on the other hand, too much noise appears to remain in the output, use -f 1.nn. You could actually go higher than 1.99 but if the scale factor is too large, you will get back the original image. When not selected, the scale factor defaults to 1.0.
- r The floating point number following -r tells *mclean* to amplify the thresholded residual, RT, before recombining it with the smoothed image, S. This exaggerates the intensity of the remaining features. If the number is more than about 1.05, the result looks artificial. This defaults to 1. When not selected, RT is not amplified.
- h Flag -h tells *mclean* to use an alternate method to select the lower and upper thresholds. When -h is NOT present (the default) *mclean* computes the thresholds as follows: Compute a graylevel histograms ht and hb from the tophat image T and bothat image B, respectively. ht is defined over nonnegative values only and hb is defined over nonpositive values only. *mclean* takes as lower threshold, l, the second moment of hb. Similarly, u is the second moment of ht. When -h IS present, *mclean* creates one histogram, h defined over all values. It chooses upper and lower thresholds as the value one standard deviation above and below the mean, respectively. Generally the default procedure does a better job because in most images there is an unequal distribution of light and dark features (and noise).
- a Switch -a tells *mclean* to write out the smoothed image to a file where OCfile represents the user-supplied file name. It can be quite useful to have this, sometimes, for experimentation.
- k Switch -k indicates that the next field defines the morphological structuring element (SE) in exactly the same way as for program *morph*.



#### IV. 3D IMAGE MORPHOLOGY SOFTWARE

This section details *Morph3DSub*, the 3D morphology subroutine, and *morph3d* the user program for 3D morphology that operates on sequences of Sun rasterfiles.

##### Morph3DSub

Subroutine *Morph3DSub* performs many (if not all) possible 3D morphological operations on gray-level or binary image time sequences or 3D voxel images. These include hit-or-miss transforms, order statistic filters, function processing or set processing erosion, dilation, opening and closing, tophat (image minus opening), bothat (closing minus image), and isolated delete functions.

To use make the following call in a c program:

```
MorphSub(  MOp, InputImage, InStruct, OutputImage, OutStruct, Ix, Iy, ImgType, NZPad, LThresh, UThre
           SE, AutoSE, sX, sY, sZ, sV, sorgx, sorgy, sorgz, SEType, SorF, Rank )
```

where the parameters are defined by:

int MOp;	Morphological operation to perform
int (*InputImage)();	address of image input function
int * InStruct;	address of structure to pass to InputImage
int (*OutputImage)();	address of image output function
int * OutStruct;	address of structure to pass to OutputImage
int Ix, Iy;	image horizontal, vertical dimensions
int ImgType;	image type (gray-level or binary)
int NZPad ;	flag. F $\Rightarrow$ zeropadding of input
int LThresh;	lower binary threshold value
int UThresh;	upper binary threshold value
int * SE;	SE in plane-row-major order
int sX,sY,sZ,sV;	SE x, y, and z support dims and max gray-lev (v)
int sorgx,sorgy,sorgz;	SE origin coordinates
int SEType;	Binary SE or gray-level SE
int SorF;	Set operation or Function operation
int Rank;	rank for rank filter

A detailed explanation of parameters follows:

Mop Morphological Operation. This integer parameter can have the following values:

mnemon	hex	dec	action
ERODE	0x0001	1	erosion
DILATE	0x0002	2	dilation
OPEN	0x0004	4	opening
CLOSE	0x0008	8	closing
RANK	0x0010	16	rank filter (order statistic filter)

<b>MAXMIN</b>	<b>0x0020</b>	<b>32</b>	<b>max of minima from nbhd. in each image</b>
<b>MINMAX</b>	<b>0x0040</b>	<b>64</b>	<b>min of maxima from nbhd. in each image</b>

Moreover, the erode and dilate functions can be "negated" using:

<b>mnemon</b>	<b>hex</b>	<b>dec</b>	<b>action</b>
<b>NOTFLG</b>	<b>0x0080</b>	<b>128</b>	<b>"not" flag for binary erodes and dilates</b>

Let I = original image; E = eroded image; D = dilated image. The following are valid parameters:

<b>mnemon</b>	<b>hex</b>	<b>dec</b>	<b>action</b>
<b>ERODE   NOTFLG</b>	<b>0x0081</b>	<b>129</b>	<b>I &amp;&amp; !E (pixelwise)</b>
<b>DILATE   NOTFLG</b>	<b>0x0082</b>	<b>130</b>	<b>D &amp;&amp; !I (pixelwise)</b>

**ERODE | NOTFLG** with a binary hit-or-miss structuring element will delete in a binary image sequence, white features with the shape of the "hit" portion of the SE. (e.g. one can easily devise a SE to delete isolated white pixels). With a gray-level SE it will delete the "interiors" from sets of white pixels in a binary sequence.

**DILATE | NOTFLG** with a binary hit-or-miss structuring element will delete in a binary image seq., black features with the shape of the "hit" portion of the SE. (e.g. one can easily devise a SE to delete isolated black pixels). With a gray-level SE it will delete the "interiors" from sets of black pixels in a binary image sequence.

Each of these operations occurs over a 3D data set within a 3D neighborhood (nbhd) defined by the 3D shape of the structuring element (SE). The definitions are exactly the same as their 2D counterparts. (Mathematical morphology is defined in n dimensions.) The exceptions are maxmin and minmax which have no counterparts in 2D. These operators were designed for use with time sequences. Consider the 3D SE to be a stack of z 2D SE's. The 2D SE's trace nbhds in adjacent sequential images. Maxmin finds the min in each 2D nbhd and takes the maximum of these. Minmax takes the smallest maximum from among the 2D nbhds. The idea behind minmax is this: Applied to a time sequence, a 3D structuring element demarcates an area (a nbhd) in successive images. (The nbhd in an image may or may not be the same the nbhds in successive images.) If a bright, moving particle stays within the successive nbhds the result of the minmax will be a bright pixel. If, on the other hand, the particle moves out of one of the nbhds, the result will be a darker pixel. Thus, when used with minmax, an appropriately shaped SE can act as a velocity filter. Maxmin is the same with respect to dark particles.

**I/O** InputImage(), InStruct, OutputImage(), OutStruct

These are pointers to user supplied IO routines and data structures that Morph3DSub uses to read and write images. These are described below.

**Ix** Image horizontal dimension. Number of pixels per line (or number of columns per row) in an image from the sequence.

**Iy** Image vertical dimension. Number of lines (or rows) in an image.

All images in the input sequence must have the same xy dimensions.

**ImgType** Image type: This specifies whether the input image sequence is to be treated like a gray level or binary sequence. The values in a binary image are zero and not-zero (it does not matter which of the values in [BLACK,WHITE]). Parameter values:

mnemon	hex	dec	action
GRAIMG	0x0000	0	gray-level image sequence
BINIMG	0x0200	512	binary image sequence

**NZPad** If zero (FALSE) Morph3DSub will extend the size of each input image internally and create a border of zeros around it in the *x* and *y* dimensions. If nonzero (TRUE) Morph3DSub will not do the zero padding.

There are image border effects with each NZPAD option. The effects differ depending on the option. With the option FALSE, Morph3DSub does what it can to "color in" each image near its perimeter. With this option TRUE, each output image has a border of zeros inside it the width and height of the SE. That is, the transformed area of each output image is smaller than the actual image dimensions. This is, in a sense, a more accurate result than the zero padded default. To zero pad the input permits the program to transform the border region, but it does this on the assumption that the original scene was black outside the image sequence. This, of course, is almost never true. Thus, the border region is inaccurately transformed. Use this switch if accuracy is more important than having an image sequence that is "colored in" out to the boundary.

**Notes:** This is all internal to Morph3DSub. The input and output images are all Ix by Iy independent of this flag.

Morph3DSub *always* zero pads in the *z*-dimension. That is it always catenates enough zero images to the beginning and end of the sequence so that the number of images output equals the number of images input.

**LThresh** lower binary threshold.

**UTHresh** upper binary threshold.

These two parameters will extract binary images from grayscale images via thresholding if: UThresh is non zero AND SorF is SET. Then all pixels with grey-levels in [Lthresh,Uthresh] will be set to WHITE, all others to BLACK. Thresholding occurs before any morphology. This makes it possible to specify binary morphology on a graylevel image sequence. If UThresh > 0 AND LThresh > Uthresh, Morph3DSub aborts with an error. If UThresh is zero no thresholding is performed. Note that one can use Morph3DSub to perform a simple threshold on an image by specifying MOp = ERODE, ImgType = GRAIMG, AutoSE = AUTO sX = 1, sY = 1, sZ = 1, sV=1 and the appropriate LThresh and UThresh.

**SE** SE, sX, sY, sZ, sV, sorgx, sorgy, sorgz

This is a pointer to the structuring element (SE). Conceptually, the SE fits in a rectangular box with dimensions sX, sY, sZ in 3-space. Morph3DSub treats the SE like it is a collection of sZ rectangles, each with sX elements in the horizontal dimension and sY in the vertical. The origin of the SE is specified by (sorgx,sorgy,sorgz) where (0,0,0) refers to the front (i.e. first rectangle) upper lefthand corner of the box. The origin must lie within the enclosing box or Morph3DSub aborts. The SE is stored as a list of ints in plane-row-major order. That is the first int is the front upper lefthand corner of the SE and the last int is the back lower righthand corner of the SE. Sequentially through the int list is the first row of the first plane of the SE, followed by the second row of the first plane, etc., until the first plane is specified, then comes the second plane in row-major order, and so on.

sZ is a particularly important parameter. Morph3DSub uses this value as the z depth of its 3D image buffer.

The 3D SE scans the image sequence in 3D moving window fashion. That is, the SE moves systematically over the entire 3D image so that the origin of the SE coincides with each pixel in the image exactly once. The way I have set up the program, for each position of the moving window scan, consecutive planes of the 3D SE lie in consecutive 2D images from the sequence. To wit, SE planes and image (voxel) planes coincide. At any point in the scan, the output image is in the same sequential position as the image that contains the SE origin.

All seven of the above parameters can be supplied to Morph3DSub by the companion routine GetSE(). The functionality of GetSE was made separate from MorphSub3D because a calling program may need to know the structure of the SE (in particular its z-dimension size) to set up image IO. (E.g., so it knows which output image corresponds to which input image).

A more complete description of the structure of the SE is given below.

**SEType** specifies whether the structuring element is to be interpreted as a gray-level SE or a binary SE. (See below for the distinction.) Its possible values are:

mnemon	hex	dec	action
GRASE	0x0000	0	gray-level SE
BINSE	0x0100	256	binary SE

**SorF** Specifies whether the morphological operation is to be of the "set" type or "function" type. (See below for the distinction.) Its possible values are:

mnemon	hex	dec	action
SET	0x0000	0	set operation
FUNCT	0x0400	1024	function operation

**Rank** This is meaningful only if Mop == RANK. Then the parameter indicates the order of the filter. The rank option of Morph3DSub is actually an order statistic filter. To compute

the order, Morph3DSub counts the number of pixels in the 3D support of the SE. In a gray-level SE that is all values greater than or equal to zero. We did not define a hit-or-miss rank filter. Therefore, to compute the support of a binary SE, Morph3DSub changes the zeros in a binary SE into -1's (DON'T CARES). Then all the rank routine considers the support to be all SE elements  $\neq$  ZERO.

If Rank == 0, Morph3DSub applies a median filter. If Rank == 1 it does a dilation. If Rank == support of SE, then Morph3DSub does an erosion. Otherwise, Morph3DSub does an order statistic filter of order Rank.

User supplied I/O routines:

Routine Morph3DSub operates on a 3D image as a sequence of 2D images. The sequence can be either a time sequence of 2D images or "plane slices" of voxels. Because there are few workstations that can load a large 3D image into memory all at once, Morph3DSub scrolls through the sequence keeping only the minimum number in memory at once. (The minimum number is SZ, the z-dimension of the SE.)

I wanted Morph3DSub to be independent of any particular image file format. This meant that Morph3DSub could not easily do its own file I/O. However, because of the size of 3D images, the 3D morphology program needs to read images successively from files. Thus, I designed Morph3DSub to "call back" to its host program for the next image in the input sequence. Likewise, it calls back to output an image. This approach lets a user write a host program to do file I/O on any format he likes and call Morph3DSub to operate on them. To do this, a user needs to write, along with the host program, an image input routine and an image output routine. The user passes the addresses of these two "call back procedures" to Morph3DSub during its function call. If ImageIn() is the input procedure and ImageOut() the output procedure that the user has written, Then the host program calls Morph3DSub() like this:

Morph3DSub( MorphOp, ImageIn, InParams, ImageOut, OutParams, ...

InParams and OutParams are pointers to structures that contain all the data necessary for ImageIn() and ImageOut() to work, yet about which Morph3DSub() does not need to know.

Morph3DSub receives the address of the image input routine in variable InputImage. It gets the address of the output routine in OutputImage. The local input and output data structure pointers are passed in InStruct and OutStruct, respectively. Whenever Morph3DSub needs a new image, it calls

IOError = InputImage( In, NumInReqs, &eof, InStruct );

Whenever Morph3DSub is ready to output an image, it calls

IOError = OutputImage( Out, NumOutReqs, OutStruct );

Morph3DSub expects the call back procedures to have the following structure:

```
int ImageIn( In, NumReqs, eof, InStruct )
    byte *In;
    int NumReqs;
    int *eof;
    struct IOS *InStruct;
    {
        1. Figure out name of next input file;
        2. Read next image file and set *eof = FALSE;
           (If there was an error in the read, return with error.)
        3. If the file could not be read because there are no more
           images set *eof = TRUE and return with no error;
        4. Decode image and extract luminance information from image
           (lumi. extraction not necessary if image is grayscale);
        5. Copy grayscale image in row major order, one byte per pixel,
           to consecutive locations starting at In;
        6. Return with no error.
    }
```

and

```
int ImageOut( Out, NumReqs, OutStruct )
    byte *Out;
    int NumReqs;
    struct IOS *OutStruct;
    {
        1. Figure out name of next output file;

        2. Copy grayscale image in row major order, one byte per pixel,
           from consecutive locations starting at Out;
        3. Construct output image in appropriate format;

        4. Write image to file;

        5. If write was successful, return with no error.
           (Else return with error.)
    }
```

The names of the two callback procedures are defined by the user. For simplicity of description, I will assume they have the names ImageIn() and ImageOut().

The first argument, In, of the ImageIn is a pointer to an image array. Morph3DSub allocates this array and passes its address to the ImageIn through this variable. ImageIn reads and decodes an image file. It extracts the luminance component and writes the gray-level pixels (one byte per pixel) in row-major order starting at the address in In.

NumReqs is the number of input requests that have been made by Morph3DSub. When

**ImageIn** is called the first time **NumReqs == 1**. **ImageIn**, may or may not use this information.

The third argument, **eof**, is a pointer to a flag in **Morph3DSub**. **ImageIn** must write a zero there whenever it inputs an image. If there are no more images in the input sequence, **ImageIn** must write a nonzero number there.

**InStruct** is a pointer to a structure that includes information necessary for **ImageIn** to work but is not directly needed by **Morph3DSub**. **InStruct** is **ImageIn**'s link back to the host program. For example, **InStruct** may contain image header and colormap information, filename tags, or information to be conveyed to **ImageOut**.

The first argument, **Out**, of the **ImageOut** is a pointer to an image array. **Morph3DSub** allocates this array and passes its address to the **ImageOut** through this variable. **ImageOut** creates an output file incorporating the image in this array.

**NumReqs** is the number of output requests that have been made by **Morph3DSub**. When **ImageOut** is called the first time **NumReqs == 1**. The output routine, **ImageOut**, may or may not use this information.

**OutStruct** is a pointer to a structure that includes information necessary for **ImageOut** to work but is not directly needed by **Morph3DSub**. **OutStruct** is **ImageOut**'s link back to the host program. For example, **OutStruct** may contain image header and colormap information, filename tags, or information to be conveyed to **ImageIn**. It can be convenient to have **OutStruct == InStruct**.

Both callback procedures return a functional value. A zero returned indicates normal completion of the routine. A nonzero value indicates an I/O error that should abort the entire operation.

If you want to write a host program for **Morph3DSub**, I highly recommend that you look at the example I/O routines **InputImage()** and **OutputImage()** in **morph3d.c**. **morph3d.c** is a sun rasterfile based host program for **Morph3DSub**.

Structuring element specifications:

A companion routine, **GetSE**, loads or creates structuring elements for **Morph3DSub**. It returns a pointer to the SE list. Its calling sequence is defined as follows:

```
int *GetSE( SENAME, AutoSE, SEType, SorF, sX, sY, sZ, sV, sorgx, sorgy, sorgz )
    char *SEName;          filename of SE file or NULL
    int AutoSE;             nonzero  $\Rightarrow$  create SE internally
    int *SEType;            binary or graylevel
    int *SorF;              set or function
    int *sX,*sY,*sZ,*sV;    encl box size and max gray level
    int *sorgx,*sorgy,*sorgz; origin cdt.
    {
    body of routine
    }
```

**SEName:** This is the filename of a SE file. If the environment variable SEPATH is defined, the value of it is prepended to the filename prior to opening the file. If this value is NULL, GetSE() assumes that it is being asked to roll its own SE.

A structuring element file is an ASCII file of integers separated by spaces. The first three numbers, x, y, z are the horizontal, vertical, and depth dimensions in pixels of the smallest 3D box that will cover the structuring element. Numbers x, y and z must be > 0. The next three numbers, i, j, k are the horizontal, vertical, and depth coordinates, respectively, of the SE origin. **IMPORTANT:** The origin is expected to be in the covering box. If not, MorphSub3D aborts. The front upper left hand corner of the box has coordinates (0,0,0); the back lower right is (x-1,y-1,z-1). Following the first six integers are x\*y\*z integers separated by spaces. These numbers are the SE elements. Their interpretation depends on the morphological operation being performed.

Negative SE elements are ALWAYS treated as logical DON'T CAREs. That is, when the operation is in progress, image pixels under negative SE elements are ignored. Thus, the support of the SE is limited to those elements that are nonnegative. This permits the creation of odd-shaped and multiply connected SE's or the placement of the SE origin outside the body of the SE. If ImgType is binary, (i.e. pixels grouped as zero and not zero), and if SEType is binary, then the SE is used to perform a hit-or-miss transform. In this case, zero SE elements cover the "miss" support and positive (nonzero) elements cover the "hit" support. The actual gray-levels are ignored.

If ImgType is binary, and SEType is gray then the nonnegative (both zero and greater than zero) SE elements determine the support of a "hit-only" transform. That is, the nonnegative support is used as a standard set-type SE for set (binary) morphology. (Of course, the other gray-level info is ignored.) Note: if ImgType is binary, then a set operation is performed by default (SorF is ignored).

The interpretation of the SE elements for ImgType gray depends on the flags SEType and SorF, just as for *morph*:

**AutoSE:** If this value is zero an SE file is read in. If it is nonzero, it tells GetSE() to either use one of the predefined (canned) SEs or to create an SE. Here are the possible values for AutoSE and their implications.

mnemon	dec	meaning
PLUS	1	use 3 by 3 by 3 3D "+" shaped SE
S3X3X3	2	use 3 by 3 by 3 cube SE
S5X5X5	3	use 5 by 5 by 5 quasi disk SE
CONE	4	make a conical SE
CYLINDER	5	make cylindrical SE
SPHERE	6	make spherical SE

The first three SEs are canned SEs. That is they exist as data structures in the compiled code of Morph3DSub. The other three SEs are built by Morph3DSub to user specification. When any of the six are requested SEType is set to gray (SEType == GRASE). If one of the canned SEs is chosen a set operation is performed (SorF == SET) overriding the input parameter. If CONE, CYLINDER, or SPHERE is chosen a set operation is performed if sV == 0, and a function operation is performed if sV != 0.



**PLUS:** This SE has a voxel at its origin and another attached to each face of the origin voxel for as total of 7 voxel elements. It looks like a "+" in orthographic projection from any x, y or z, direction.

**S3X3X3:** This is simply a 3x3x3 cube of voxels with origin at the center.

**S5X5X5:** Extend each face of the 3x3x3 cube outward by one layer of voxels and you get this SE. It looks like a fat, short-armed 3D "+".

**CONE:** GetSE makes an elliptical cone as follows: The origin is a single voxel at the center of the SE. The axis of the cone is parallel the z-axis (it is perpendicular to the image/voxel planes). Each end of the cone is an  $sX$  by  $sY$  elliptical disk (circular disk, if  $sX = sY$ ).  $sX$ ,  $sY$ , and  $sZ$  must be odd and greater than or equal to 1. Each end of the cone is  $(sZ-1)/2$  image planes away from the plane of the origin. For example, if  $sZ = 3$ , the ends of the cone are in the planes adjacent to the origin plane. If  $sZ=5$  there is one image plane between the end plane and the origin plane at each end. The slope of the cone in the  $xz$  plane is  $sX/sZ$  and in the  $yz$  plane the slope is  $sY/sZ$ . SE planes between the origin and the ends are elliptical disks with semimajor and semiminor axes determined by the slopes. The cone is actually an hourglass in shape. This shape can track pixels in a time sequence that have a maximum velocity in  $(x,y)$  given by  $(sX/sZ, sY/sZ)$ . If a graylevel value,  $sV > 0$ , is requested, the voxels along the cone axis are given value  $sV$ . The graylevels assigned to other voxels decrease as the square root of the radius to zero at the edges of the cone.

**CYLINDER:** GetSE makes an elliptical cylinder as follows: First, it makes an elliptical disk with a diameter of  $sX$  pixels horizontally and  $sY$  pixels vertically.  $sX$  and  $sY$  must be odd and greater than or equal to 1.  $sV$  is the gray level of the center voxel. If  $sV > 0$  is requested the gray level of the voxels decrease as the square root of the radius to zero at the edge of the disk. Then, the disk is copied  $sZ$  times. The origin of the SE is placed at its very center (center of the disk at  $sZ/2$ ). A cylindrical SE can track shapes in a time sequence larger than its own cross-section that are not moving over  $sZ$  frames. A cylinder of diameter 1 can track stationary pixels.

**SPHERE:** This is a spheroid of dimensions  $sX$  by  $sY$  by  $sZ$ . It is made by GetSE in a fashion analogous to the construction of a single identical to one slice of the cylinder. If  $sV > 0$   $sV$  is the value of the center pixel and the gray-values decrease as the square root of the radius to zero at the edges. The graylevel sphere will approximate densities in a 3D voxel image.

Note that when  $AutoSE \neq 0$ , the values of  $SEType$  and  $SorF$  are ignored.

## References

- [1] Giardina, C. R. and E. R. Dougherty, *Morphological Methods in Image and Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1988.
- [2] Haralick, R. M., S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology" *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, No. 4, pp. 532-550, 1987.
- [3] Maragos, P. and R. W. Schafer, "Morphological filters - part I: their set theoretic analysis and relations to linear shift invariant filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, No. 8, pp. 1153-1169, 1987.
- [4] Maragos, P. and R. W. Schafer, "Morphological filters - part II: their relations to median, order-statistic, and stack filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, No. 8, pp. 1153-1169, 1987.
- [5] Maragos, P. and R. W. Schafer, "Morphological Systems for multidimensional signal processing," *Proc. IEEE*, vol. 78, No. 4, pp. 690-710, 1990.
- [6] Matheron, G., *Random Sets and Integral Geometry*, Wiley, New York, 1975.
- [7] Ronse, C., *Why Mathematical Morphology Needs Complete Lattices*, Manuscript M-270, Philips Research Laboratory Brussels, Avenue Van Becelaere 2, Box 8, B-1170 Brussels, Belgium, November, 1988.
- [8] Serra, J., *Image Analysis and Mathematical Morphology*, Academic Press, London, 1982.
- [9] Serra, J., "Introduction to mathematical morphology," *Comp. Vision, Graph., Image Process.*, vol. 35, pp. 283-305, 1986.
- [10] Serra, J., ed., *Image Analysis and Mathematical Morphology, Vol. 2: Theoretical Advances*, Academic Press, New York, 1988.
- [11] Sternberg, S. R., "Grayscale morphology," , vol. 35, *Comp. Vision, Graph., Image Process.*, pp. 333-355, 1986.

## **A REVIEW OF CADDMAS**

**Dean Lance Smith, Ph.D., P.E.  
Associate Professor  
Department of Electrical Engineering  
Memphis State University  
Memphis, TN 38152**

### **Abstract**

The CADDMAS project was reviewed. The architecture of the prototype of the proposed full scale system appears sound, although some potential problems will probably be found as the users of the system gain more experience. The software development and operating environment for CADDMAS is adequate, but less than ideal. The ideal software environment will probably not exist for several years until better software standards are developed for parallel computer systems.

### **Introduction**

The CADDMAS project at Arnold Engineering Development Center was reviewed during the summer of 1991 under the sponsorship of the United States Air Force, Office of Scientific Research, Summer Faculty Research Program. Mr. Tom Tibbals of Sverdrup Technology, Inc. was the focal point for the program. The review covered three aspects of the CADDMAS project: 1) the appropriateness of the proposed architecture, 2) the potential for

running existing software on CADDMAS, and 3) the potential for assigning the CADDMAS resources dynamically.

### **Description of CADDMAS**

CADDMAS is an acronym for Computer Assisted Dynamic Data Monitoring and Analysis System. The proposed system will acquire data in digital form from sensors mounted on experimental turbine engines while the engines are tested. CADDMAS will also analyze the data, and report the results in real-time. Details of a scaled down, proof-of-concept system have been reported elsewhere [1].

CADDMAS is a multi-processor computer system. Several types of processor elements are used in the system. Most of these processing elements are replicated through out the system.

One processing element is a dedicated data acquisition system based on the Motorola DSP56001 DSP microprocessor. A Motorola DSP96002 DSP microprocessor will be used in the proposed system. The system acquires data from a turbine sensor, does some analysis on the data, and reports the results to the rest of CADDMAS. The proposed system will also have built-in self-test capability.

Another processing element is based on the INMOS Transputer [2]. The Transputer is a family of microprocessors that can be used to construct MIMD (Multiple Instruction stream, Multiple Data stream) parallel computer systems. Several of the Transputer elements have Zoran ZR34325 vector signal processors as co-processors [3].

Data can be saved on one or more data storage elements. These systems can save data on either hard disk or cassette tape.

The user of the system communicates with the processing elements through a personal computer. Results can be displayed on the personal computer's monitor. However, the system also has several high performance graphic monitors for displaying graphs [1].

All of these processing elements are connected together with serial communication links based on the Transputer protocol. Some of these links are static. Others are dynamic and can be altered by writing the appropriate bit pattern into an INMOS IMS C0xx family Programmable Link Switch [2].

CADDMAS will replace a current system where data is recorded on analog tape. The analog tape data is converted to digital code off line and analyzed on mainframe computers or vector supercomputers.

## **The CADDMAS Architecture**

The review of the architecture was based on reading some of the literature cited in this report and some internal reports that were not cited. The review was also based on several discussions with Mr. Tom Tibbals of Sverdrup, and Mr. Ted Bapty and Mr. Ben Abbott of Vanderbilt University, the engineers who have done most of the design and development work on CADDMAS. Discussions were also held with Mr. Jim Nichols of Sverdrup, an engineer who will probably use the system. Based on the information available to the author, it appears the proposed architecture of CADDMAS is well thought out and that the engineers primarily involved in the project are well aware of some of the potential problems and risks they face developing a final system that will meet the system's specifications.

Several strategies have been used to obtain the high processing rates required for CADDMAS. The multi-processor and parallel processor approach using relatively inexpensive microprocessor components is cost effective. Using dedicated processors for data acquisition and analysis of some of the data that will be required by all users of CADDMAS is common practice for large, dedicated data acquisition systems. This approach is indeed appropriate for CADDMAS. The dedicated processors are cost effective and will free up the remainder of the system for user specific data analysis. The design engineers appear to be willing to add more dedicated processor capability if, and when, data processing functions common to most CADDMAS

users are identified. This philosophy is appropriate, clearly within mainstream computer engineering thought, and strongly encouraged by the author of this study.

The use of vector co-processors in one or more Transputer elements is also appropriate and advantageous. The cost of adding a high performance integrated circuit vector processor is relatively small. Vector processors have definite advantages for handling fine grain parallel processing algorithms, such as some types of array processing. A MIMD architecture is more appropriate for handling medium and coarse grain parallel algorithms, such as processing similar data from several processor channels. Some of the engineers who will use CADDMAS have already had experience running software on vector processors, such as the Cray supercomputers. The vector co-processors in some Transputer elements will provide a vehicle for expressing fine grain parallelism.

MIMD parallel processing using low cost integrated circuit technology is clearly an appropriate design strategy for the CADDMAS project because of the project's high data processing rate requirements. The extensibility of the design is also desirable since the demand for data processing capability will probably expand as the users of CADDMAS gain more experience with multi-processor data processing. The design engineers are well aware that the major problem with using a parallel processing architecture, be it vector processing or MIMD, is writing effective and efficient software for the

systems. Some of the issues involved are discussed in a later section of this report. However, two trends in the development of technology are minimizing this risk.

One trend is that the performance/price ratio of semiconductor integrated circuit technology is halving approximately every 18 months. This trend will probably continue for the immediate future. As a result of this trend, even the inefficient use of the processing power of MIMD hardware is becoming more and more cost effective.

The second favorable technology trend is that several novel approaches to more effective software development for MIMD systems are being explored. The low cost of microprocessor-based MIMD architectures has generated a lot of research into effective programming techniques for these system. Some of these approaches are discussed later.

While some risk is involved in using a parallel architecture, it is a reasonable risk given the low hardware costs associated with MIMD design. Moreover, improving technology is decreasing the risk.

The choice of the Transputer for the main processing element is also an appropriate choice. There are several microprocessor chips that could be used in a MIMD system. Intel uses its own 80386 and 860 chips in the MIMD systems



it sells through its Intel Supercomputer Systems Division. The Transputer has several features that make it attractive for MIMD systems [2].

A linear arrangement of Transputers is probably appropriate for most of the data processing of CADDMAS since this arrangement corresponds to the physical arrangement of the sensors on a turbine engine. However, the flexibility built into the connections of the Transputers in CADDMAS is clearly appropriate. The Programmable Link Switch is an appropriate strategy for quickly moving sensor data to a graphical display. Some users of the system may wish to cross correlate data between sensors. The Programmable Link Switch in CADDMAS gives the user the ability to exchange data between processors directly, rather than transmitting the data through another processor. This feature will improve performance if used properly. Exactly which Transputer elements will need to be connected through the Programmable Link Switch will probably need to be settled after the CADDMAS users have gained experience with the system.

Using an IBM PC style personal computer as a console for the system is also appropriate. A workstation would provide a higher performance user interface with superior graphics capability. Unfortunately, most of the CADDMAS users are probably unfamiliar with workstations and the UNIX operating system that runs on most workstations. The users are more likely to be comfortable with an MS DOS, PC style interface. They may also wish to run PC programs such as Lotus 1-2-3. The selection of a PC interface should not

seriously degrade system performance and will probably improve its utility. CADDMAS has high performance graphics monitors as part of its architecture. Most users will probably not need the higher processing performance of a workstation since there should be adequate capability from the Transputer elements. Further, the performance of PCs is approaching that of workstations as technology improves. This trend is minimizing the performance difference between PCs and workstations.

The engineers developing CADDMAS are concerned about the compatibility of CADDMAS software with new processing element technology. This concern is indeed valid. Most of these issues are covered in a later section of this report. It is important to note that INMOS has introduced higher performance versions of the Transputer chips as semiconductor technology has improved [2]. The improved chips have machine code compatibility with the older chips. Other compatibility issues are discussed in a later section.

There is one potential problem that AEDC may have with CADDMAS. The DFT is based on several assumptions about the vibrations, the most important of which is that the vibrations are sinusoidal [4], [5]. Sinusoidal vibrations always occur in homogeneous materials with rectangular boundaries (for two dimensions) or plane boundaries (for three dimensions) [6]. Any curvature in a boundary will tend to focus the spacial waves and produce standing waves in space. The time waveform at any point in space

will still be sinusoidal. However, the spacial waveform will not be exactly sinusoidal even though it may be close to sinusoidal. Using a DFT to analyze a time waveform at any point in space should introduce no problem, provided the DFT is interpreted correctly. Correlating the time waveforms from two points in space may produce problems unless the focusing effects of boundary curvature are taken into account.

The inhomogenaities and anisotropic behavior of some composite materials will introduce additional boundaries that will produce additional reflections. The additional boundaries can introduce additional harmonics that are not integer multiples of the harmonics produced by the physical boundaries of the materials even if all the spacial harmonics are sinusoidal. The inhomgenaities can also produce focusing that will tend to produce non-sinusoidal standing waves in space.

The issues involved are discussed in more detail in another publication [7]. The effects of curvilinear boundaries, inhomogeneous materials, or anisotropic materials may produce the demand for other discrete transform software that is more appropriate for the vibrations that may be encountered.

### **Running Existing Software on CADDMAS**

Any software conforming to the standards of a common, higher level programming language can potentially be re-compiled to run on a Transputer

with little or no modification. (A standard is assumed to be set by ANSI or some other recognized standards agency. A common language is assumed to be one used by engineers to program two or more different computers. Fortran, C, Ada, and Modula-2 would be examples of common, standard higher level languages.) The machine code produced by the compiler may run on only one processing element unless the compiler can spot parallelism in the program. In many cases, the software will need to be rewritten to distribute the software over two or more processing elements even if the compiler can spot some parallelism in the software.

The current support for programming parallel computer systems differs from the ideal. These differences will probably exist for many years until more research is done into how to program parallel computer systems and better standards are developed for programming languages. Some of the issues are discussed below. More information is given in [8].

### **Ideal Software Support**

It would be desirable for a compiler to take an existing program written in a standard higher level programming language and automatically identify the parallelism in the program. Indeed, some progress has been made in developing compiler algorithms that will identify parallelism [9]. Whether any commercial compiler available for the Transputer [10] takes advantage of

these techniques is unknown to the author. A compiler written at Georgia Tech is reported to detect parallelism [11].

It is unlikely that even the best compiler will ever be able to identify parallelism in a program if the programmer uses an inherently sequential algorithm. A good higher level programming language for programming parallel processors needs paradigms for identifying parallelism. Unfortunately, all current standards assume sequential processing. Most Transputer compilers for common standard higher level programming languages have extensions to the standard to support expressing parallelism. The use of these extensions will complicate recompiling the rewritten software on another vendor's compiler. It will also complicate the recompilation of the software on a standard compiler once good standards for identifying parallelism are accepted and adopted.

### **Existing Software Support**

Ada, C, Fortran, and Modula-2 compilers are all available for the Transputer. So are compilers for less common languages [10].

Based on the author's experience transporting Fortran, Ada, and C programs from one machine to another, Ada programs will present the fewest re-compilation problems, and Fortran programs will present the most problems. Ada seems to have the most machine independent paradigms, and

Fortran seems to have the most machine dependent paradigms. (Character manipulation in Fortran is a good example of a paradigm that is both machine specific and covered by a poor standard.)

Two common higher level languages, Ada and Modula-2, do have some standard mechanisms that can potentially support parallel programming. Both Ada and Modula-2 support the concept of tasks [12], [13]. Alsys has developed an Ada compiler for the Transputer that does support parallel processing with static tasks [14]. Tasks, or coprocesses as they are sometimes called, were originally intended to support concurrent programming on single processor machines. However, there are many similarities between parallel processing where messages are exchanged between the processors, and concurrent processing.

Tasks are similar to procedures or functions in the way they are written. Tasks can be called and they can return values. There is a fair amount of overhead entering and leaving tasks. Tasks, as a programming concept for parallel programming, are more suitable for medium and coarse grain parallel algorithms.

The proposed new Fortran standard, FORTRAN 90, is reported to have support for programming parallel processors [15]. The standard is reported to support the programming of vector processors. Whether the proposed standard can support medium or coarse grain parallelism remains to be seen.

No Transputer compilers that the author is aware of have extensions to support parallel processing on the vector co-processor in CADDMAS. Subprograms (functions or procedures) are available that can be called to use the co-processor. Standard C compilers support a higher level construct called pointers which can be used to directly control a memory-mapped co-processor without the overhead of subprogram calls and returns [16]. Standard Ada compilers can support direct control of memory-mapped co-processors through representation specifications [12].

### **Utilizing The Capacity of CADDMAS**

Utilizing the capacity of CADDMAS is another area where the ideal environment differs from the existing environment.

#### **Ideal Operating Environment**

In the ideal operating environment, tasks should be assigned dynamically to processing elements that are available to execute a task. Multiple tasks should be assigned to one processing element if appropriate. A task should be divided and distributed over two or more processing elements if the extra processing elements become available. A processing element that fails should be detected and taken out of service.

The dynamic assignment of tasks requires a control program, or operating system. It also requires a compiler that will produce tasks in a form acceptable to the operating system so that the tasks can be assigned dynamically. Many techniques have been proposed for generating and assigning dynamic tasks in parallel computer systems. There seems to be no common agreement on the best way to do so. It will probably be some time before widely accepted standards exist in this area. Since the issues are closely related to some of the issues involved in generating standard programming languages for parallel computer systems, standards for programming languages will probably be developed first. Some of the proposed solutions are discussed in more detail elsewhere [8].

### **Current Operating Environment**

The Transputer does have several operating systems available for it [10]. At least one, Helios, supports the dynamic assignment of tasks [17], [18].

In the most common Transputer operating environment, there is no control program per se. Each Transputer does have built into its instruction set the ability to initiate, suspend, and switch tasks. It also has the ability to exchange messages between tasks and load programs into itself through one of its communications ports after restart [2]. A communication program on a host computer loads the tasks into each Transputer on restart.



Tasks are assigned statically by the programmer in most current Transputer environments. The number of tasks and the Transputer they are assigned to is determined by the programmer. These decisions are made when the tasks are written or compiled, depending on the compiler. Tasks must be reassigned and the software must be recompiled if a Transputer is taken out of service for any reason.

### Conclusions

1. The current system for acquiring and analyzing experimental turbine engine test data is antiquated and clearly needs to be replaced with a real-time, computer-based data acquisition and analysis system. The real issue is whether CADDMAS is a sound solution.
2. The architecture of CADDMAS is clearly appropriate. The engineers designing the system are approaching the problem in a realistic and responsible manner. The computer architects are using appropriate off the shelf, state of the art technology and are designing only what cannot be purchased as a turnkey product.
3. The designers of CADDMAS are attempting to take advantage of state of the art computer architecture, especially parallel processing. They really have no choice but to use a multi-processor system to obtain the computation power needed. They have identified dedicated tasks and

used co-processors or dedicated processors for these functions. Dedicated processors and co-processors are proven design techniques. Parallel processors are being used only for high power, general purpose computation. Parallel processing is a proven hardware technique, although there is some technological risk in programming parallel processors. The designers clearly understand and are concerned about these risks.

4. Most of the programming risk of using parallel processing will be in writing or purchasing software that will take full advantage of the processing capability of CADDMAS. Some of the application and operating system software will need to be developed through trial and error as the users of CADDMAS learn more about parallel processing technology and what CADDMAS can do to help them. In the process, the users of the system will probably learn more about the problems they want to solve with CADDMAS and will better be able to express their wants and needs in terms that can be understood by computer architects and programmers. Research being conducted outside AEDC into the best software techniques for programming parallel processing systems will also have a strong and favorable impact on writing better software for CADDMAS.
5. The risk is not that CADDMAS will fail. Rather, the risks are that 1) the software will not efficiently use the hardware, 2) the time to develop

efficient software may be excessive, 3) existing software may not be as portable to CADDMAS as desired by some of its potential users, or 4) the software may not be as portable as desired to any future computing equipment that will eventually replace CADDMAS.

6. The hardware for the proposed system is so inexpensive, and should become even less expensive in the future, that the risk of inefficiently using the hardware of the system can be overcome, to a significant extent, by expanding the hardware of the system. An advantage of a good parallel processing architecture is its extensibility. The proposed CADDMAS architecture will have some extensibility.
7. In spite of the potential software problems with CADDMAS, the risk to reward ratio is so high that even the worst case risk is acceptable. Indeed, continuing to use the current data acquisition and analysis system or even replacing the current analog data acquisition system with the data acquisition co-processor developed for CADDMAS, would be riskier choices. These alternatives would continue the use of obsolete technology. The current data analysis computing facilities need to be replaced with modern, less expensive technology. There would be little, if any, reward for using the obsolete technology.
8. Several key software modules that will assist a programmer writing data analysis application programs have already been written for or adapted

to CADDMAS as part of the proof of concept system. Modules include graphics display software, including an extension to support Campbell diagrams, and a simple data acquisition and analysis program. This work should continue as additional software modules are identified. One module that should be developed is a suite of diagnostic programs so that a Transputer processing element can be tested. The test module would expedite maintaining the CADDMA's system when it fails. It would also assist in the dynamic assignment of tasks to Transputer elements if a processor element fails during an engine test.

### References

- [1] T. Bapty, B. Abbott, and J. Sztipanovits, "Real-Time Turbine Engine Data Visualization," in Transputer Research and Applications 4, D. L. Fielding, Ed. Amsterdam: IOS Press, 1990, pp. 205-214.
- [2] Anon., The Transputer Databook, 2nd ed. Bristol, UK: INMOS, 1989.
- [3] Anon., ZR34325 32-Bit Floating-Point Vector Signal Processor Engineering Data. Santa Clara, CA: Zoran Corp., May 1990.
- [4] C. S. Burrus and T. W. Parks, DFT/FFT and Convolution Algorithms, Theory and Implementation. New York: John Wiley & Sons, 1985.

- [5] R. W. Ramirez, The FFT, Fundamentals and Concepts. Englewood Cliffs, NJ: Prentice-Hall, 1985.
- [6] P. M. Morse and H. Feshbach, Methods of Theoretical Physics. New York, NY: McGraw-Hill, 1953.
- [7] D. L. Smith, "Common problems using the DFT to analyze mechanical vibrations," in preparation.
- [8] D. L. Smith, "The attraction of building parallel computer systems from inexpensive processing element, and the problems," in preparation.
- [9] J. R. Allen and K. Kennedy, "PFC: a program to convert fortran to parallel form," in Supercomputers: Design and Applications, K. Hwang Ed. Los Angeles, CA: IEEE Computer Society Press, 1984, pp. 186-203. Reprinted from The Proc. of the IBM Conf. on Parallel Comput. and Sci. Computations. Rome, Italy: IBM, 1982.
- [10] Anon., The Transputer White Pages, Software/Consultants Directory, 3rd ed. Bristol, UK: INMOS Ltd., Jan. 1990.
- [11] B. Abbott, personal communication, July 1991.

- [12] G. Booch, Software Engineering with Ada, 2nd ed. Menlo Park, CA: Benjamin/Cummings, 1987.
- [13] N. Wirth, Programming in Modula-2, 2nd ed. Berlin: Springer-Verlag, 1983.
- [14] J. G. P. Barnes, "Ada on transputer arrays," Applications of Transputers 1, L. Freeman and C. Phillips, Eds. Amsterdam: IOS Press, 1990, pp. 1-9.
- [15] P. Wallich, "FORTRAN forever, is it still the language of choice for science?," Scientific American, vol. 265, p. 112, June 1991.
- [16] B. Kernighan and D. Ritchie, The C Programming Language, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [17] Anon., Helios Developer's Notes. Somerset, U.K.: Perihelion Software Ltd.
- [18] D. Pountain, "A personal transputer," Byte, vol. 13, pp. 303-308, June 1988.

# WAKE AND PROJECTILE VELOCITY ESTIMATION

D. Mitchell Wilkes & Georges Badih Aboutanos

Vanderbilt University

Department of Electrical Engineering

Nashville, TN 37235

## 1 Abstract

Digital signal processing techniques were used to estimate the velocity of a projectile and its wake. The observed data was collected by one or multiple doppler radars in an underground ballistic range.

A classical FFT-based spectral estimation approach was used to identify the frequency characteristics of the data. These frequencies were later translated into velocities leading to accurate velocity profiles of the projectile and of the wake.

A C program was also developed to implement the spectral estimation approach and to provide an automated analysis of the data with extensive graphical display of the results. These graphical presentations include velocity profiles, 3-D surface plots, and contour plots.

## 2 Introduction

The main goal of this research was to develop an automated analysis technique for doppler radar data from a ballistic range. The doppler data is collected on a projectile followed by its wake, then it is automatically analysed to provide a graphical and numerical description of the velocity behavior of the projectile as well as the wake, which is more difficult to analyze.

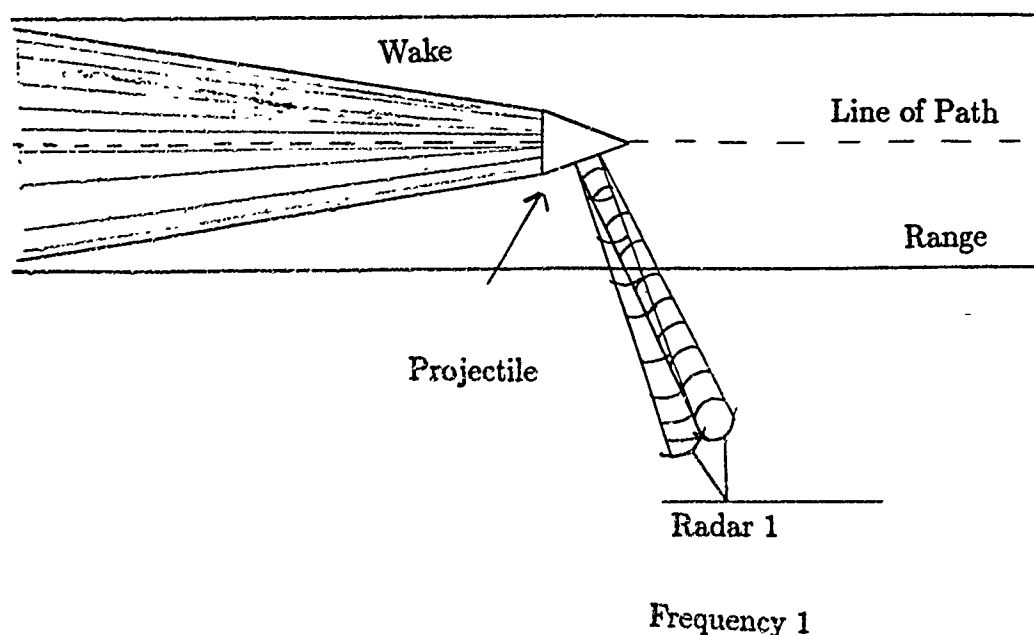


Figure 1: Projectile velocity measurement.

## 2.1 Physical Description of the Range

The measurements are taken in a long underground range, usually evacuated to a low pressure. A typical projectile is about 10 cm long and 2 cm in diameter is fired down range by a two-stage compressed light-gas gun at a speed of approximately 20,000 ft/s. As the projectile travels down the range a few doppler radars, usually three, each at a different frequency (8.6, 17 and 35 GHz), measure the velocity of the projectile and the wake following the projectile. These radar antennas are usually located about 45° from the line of path of the projectile. (See Figure 1)

## 2.2 Description of the Measurement System

This section describes the radar measurement system and the way the doppler frequency is measured and converted to velocity. The radar is transmitting at some frequency  $f_r$ . This is a quasi-monostatic radar system where the transmitter antenna and the receiver antenna are co-located. When the radar wave hits the



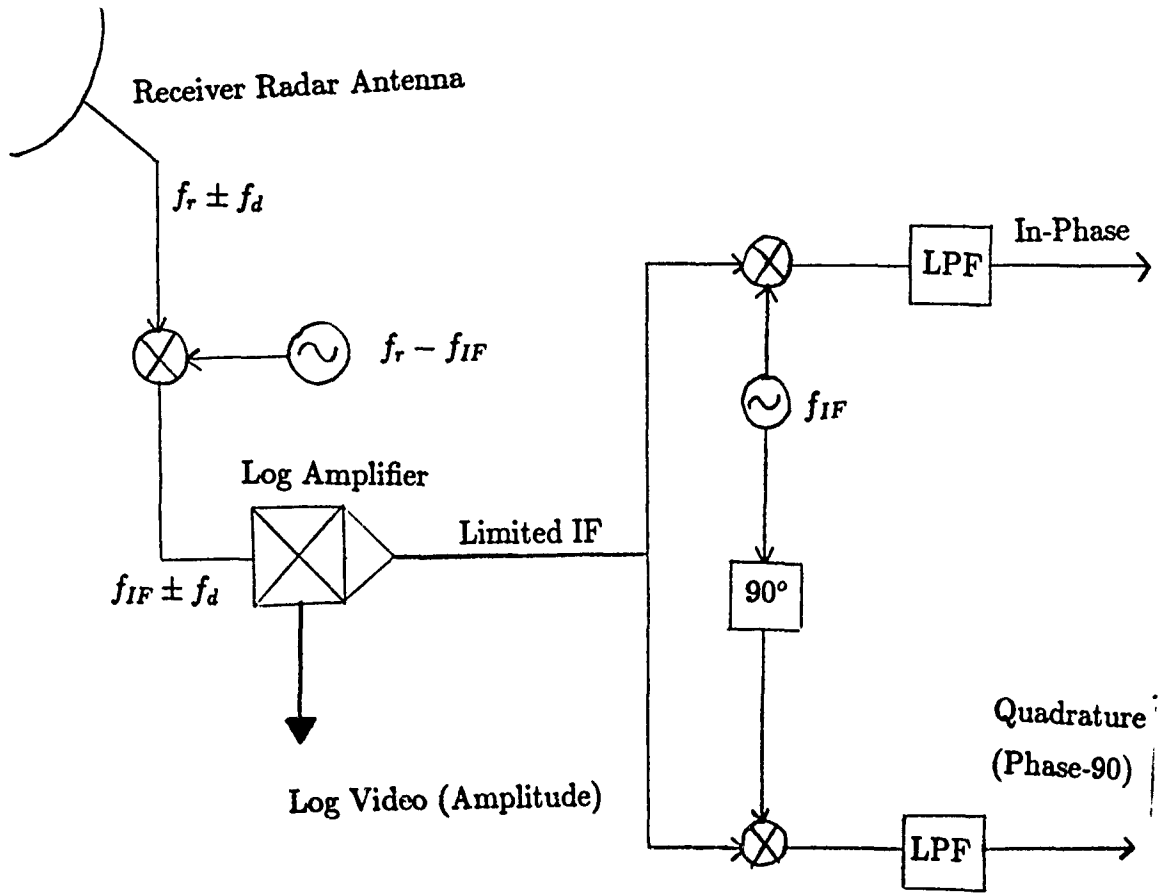


Figure 2: Radar measurement system.

moving projectile or its wake it reflects back to the radar receiver at the frequency  $f_r$  shifted up by a doppler frequency  $f_d$ .

The received frequency  $f_r + f_d$  is down converted to an intermediate frequency  $f_{IF} + f_d$ . A Log amplifier is used to collect amplitude data, and the limited IF data is separated into an In-phase channel and a Quadrature channel (which is phase shifted by  $90^\circ$  relative to the In-phase channel). The data is then downconverted to DC by multiplying it again with quadrature sinusoids at  $f_{IF}$ , and finally each channel is passed through a low pass filter leading to a doppler frequency related to the speed of the projectile and its wake at the In-phase and Quadrature channels (See Figures 2 and 3). LPF is the low pass filter,  $f_{IF}$  is the intermediate frequency,  $f_r$  is the radar frequency and  $f_d$  is the doppler frequency.

The doppler frequency is converted to velocity using the following equations [1]:

$$f_d = \frac{2v \cos(\theta) f_r}{c} \quad (1)$$

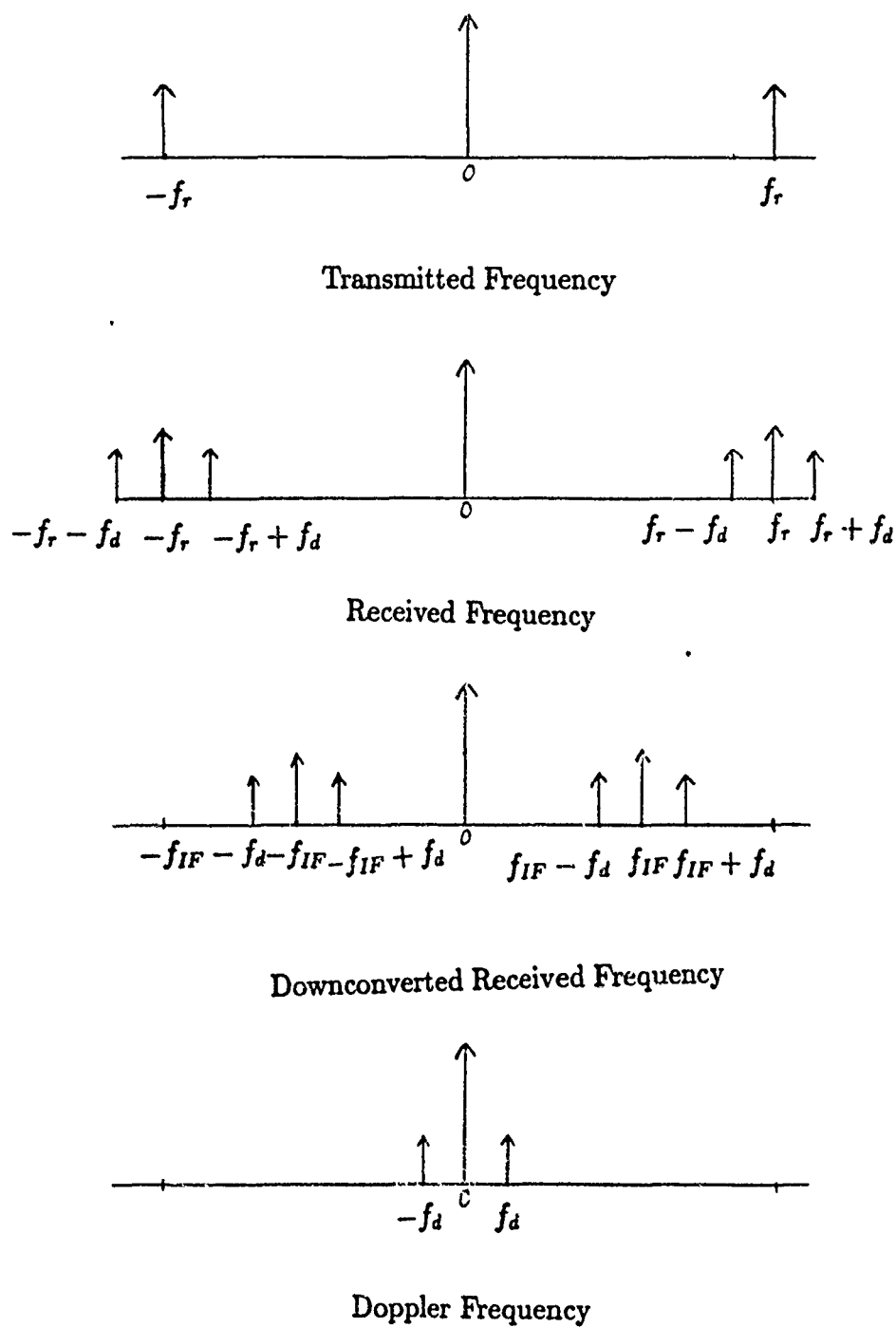


Figure 3: Radar frequency measurement.

$$v = \frac{cf_d}{2\cos(\theta)f_r} \quad (2)$$

where  $v$  is the velocity of the target,  $\theta$  is the angle between the line of the path of the projectile and the radar line of sight (usually  $45^\circ$ ),  $c$  is the speed of light,  $3 * 10^8 m/s$ ,  $f_r$  is the radar frequency, and  $f_d$  is the doppler frequency.

Substituting  $f_d$  into Equation (2) will provide the velocity of the projectile or of the wake,  $v$ . The velocity and the doppler frequency are proportional, which means the larger the doppler shift the faster the projectile is moving and the smaller the doppler shift the slower the projectile is moving.

## 2.3 Description of the Projectile and the Wake Results

The raw data is nominally taken at a sampling rate of 200 ns per point. This data typically consists of four sections. The first section is random noise at low amplitude, the second section is the return from the projectile, which is the first object of interest seen by the radar. This section is characterized by a large constant magnitude and a high doppler frequency. The third section is the return off the wake and it has basically the same constant magnitude behavior as the projectile, but the frequency is lower because the wake moves slower than the projectile. Finally the fourth section of the data is random noise again at low amplitude. An example of such data for 6000 points and a 200 ns sampling period is shown in Figure 4.

The projectile and the wake can be distinguished easily from the background noise because they have a higher amplitude. After converting the raw data to velocity we observe a null area immediately behind the projectile that is due to the recombination of gases. This feature can be used to separate the wake from the projectile so that the velocity behavior of each (i.e., projectile and wake) can be observed separately. The final result will be displayed in terms of percentage of the model speed versus distance from the beginning of the projectile measured in projectile body diameters. A sample of such a result is shown in Figure 5.

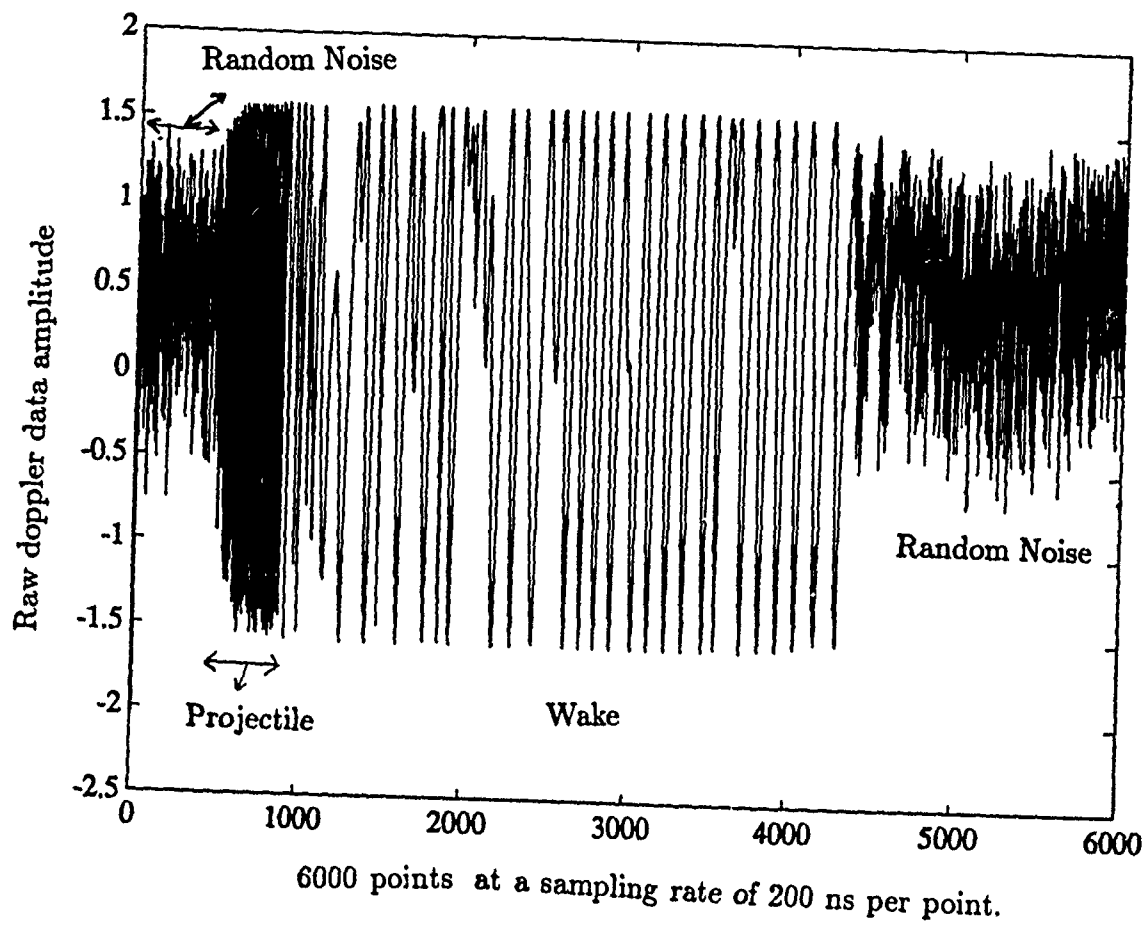


Figure 4: Raw doppler data.

## PROJECTILE+WAKE (17.0 GHz)

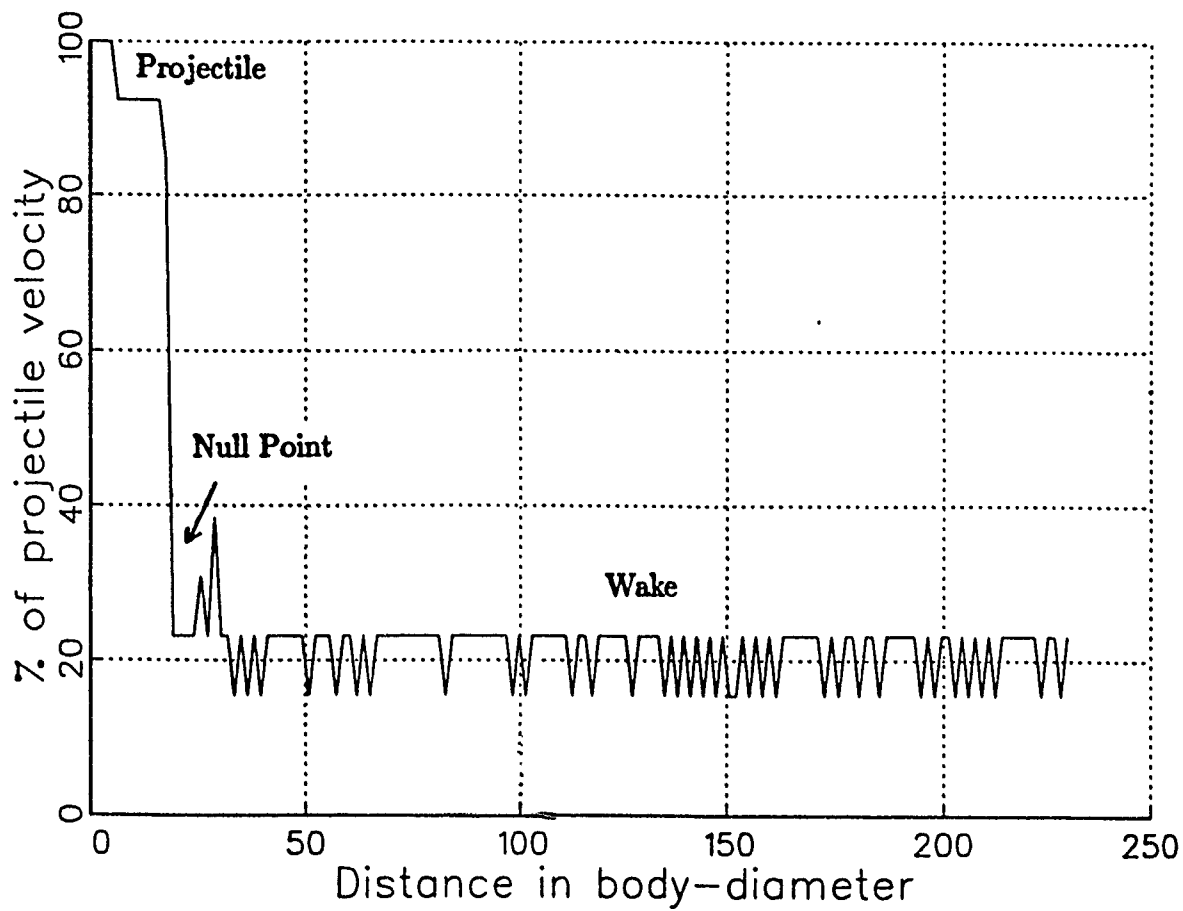


Figure 5: Extracted projectile and wake velocity profile.

## 3 Problem Discussion

### 3.1 Statement of the Problem

Provided with some doppler data collected on a projectile followed by its wake, our goal was to use digital signal processing techniques to estimate the velocity of the projectile and the velocity of the wake for possibly multiple radar frequencies such as 8.6, 17 and 35 GHz. The analysis of the data and the estimation technique should be fully automated providing a projectile and wake velocity profile, and a wake velocity profile alone. New visualization techniques were also proposed to provide a clearer understanding of the projectile and wake velocity behaviors.

### 3.2 Basic Approach

The basic approach of this effort was as follows. The doppler data is broken into many possibly overlapping windows. Spectral estimation is performed separately on each window. We find the location of the tallest spectral peak in each window and use it as our estimate of the doppler frequency for that time window. The peaks from each time window are collected and represent the doppler frequency behavior as a function of time. These doppler frequencies are then translated directly into velocities via Equation (2), and displayed in different graphical formats describing the velocity behavior of the data. Analysis is performed to extract the velocity profile of the projectile and the wake and to distinguish them from the background noise. This basic approach will now be described in greater detail.

### 3.3 Details of Spectral Estimation

Spectral estimation is used to identify the frequency characteristics of the data, which can be translated directly into velocity. A classical FFT-based spectral estimation approach was chosen due to its well behaved characteristics [2]. Although this approach may have difficulty resolving closely spaced sinusoids, this limitation does not cause a problem for a basic analysis of the data.

The first task is to break the observed sequence,  $x(n)$ , into many possibly over-

lapping  $L$  point long windows,  $x_m(n)$ , given by

$$x_m(n) = \begin{cases} x(mM + n) & n = 0, \dots, L-1 \\ 0 & \text{else} \end{cases} \quad (3)$$

where  $M$  is the distance in samples between the beginnings of successive windows and  $mM$  is the location of the beginning of the  $m^{\text{th}}$  window. Before applying FFT-based spectral estimation on each window, we applied window functions to reduce the ripples in the estimated spectrum due to sidelobe effects from using finite length data windows. Some of the windows that were used include the Hamming, Hanning, Blackman, and Rectangular windows [2]. These windows also provide a degree of spectral smoothing that tends to suppress false spectral peaks. Each  $L$  point long window is multiplied by the chosen  $L$  point long window function  $w(n)$  to form

$$y_m(n) = x_m(n)w(n) \quad (4)$$

The squared magnitude of the FFT of  $y_m(n)$  is used as the estimate of the power spectrum of the  $m^{\text{th}}$  window, given by

$$S_m(k) = |Y_m(k)|^2 = \sum_{n=0}^{N-1} y_m(n) e^{-j\frac{2\pi}{N}kn} \quad (5)$$

$N$  is the length of the FFT and should be a power of 2 in order to be able to use the common radix-2 fast algorithms [2]. We also require  $N \geq L$ , and if  $N > L$ , zeros should be appended to  $y_m(n)$  until its length is  $N$ . The frequency bin number  $k$  corresponds to a discrete-time frequency of  $\frac{2\pi}{N}k$ .

We further form a spectral matrix with columns consisting of the successive  $S_m(k)$ 's computed from the time windows. The major advantage here is that this allows us to observe all the frequencies in the data and not just the peak doppler frequencies. This provides a good 3-D view of the evolution of the doppler frequencies (i.e., velocities) over time (see Figures 6 and 7). False peaks are also suppressed by averaging the columns in the matrix three at a time to achieve some degree of temporal smoothing in the frequency domain.

# PROJECTILE+WAKE (17.0 GHz)

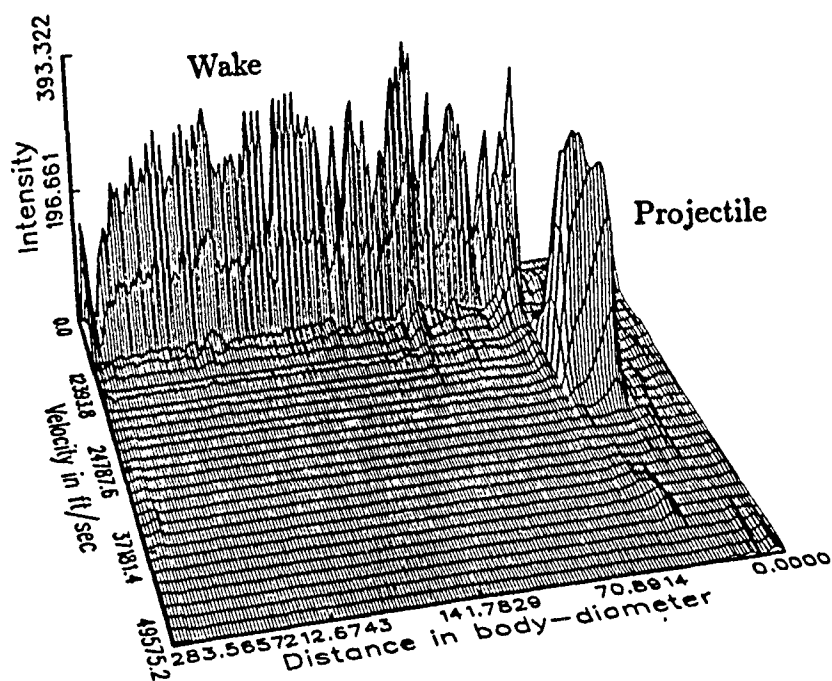


Figure 6: 3-D plot of the evolution of doppler frequency with time.



# PROJECTILE+WAKE (17.0 GHz)

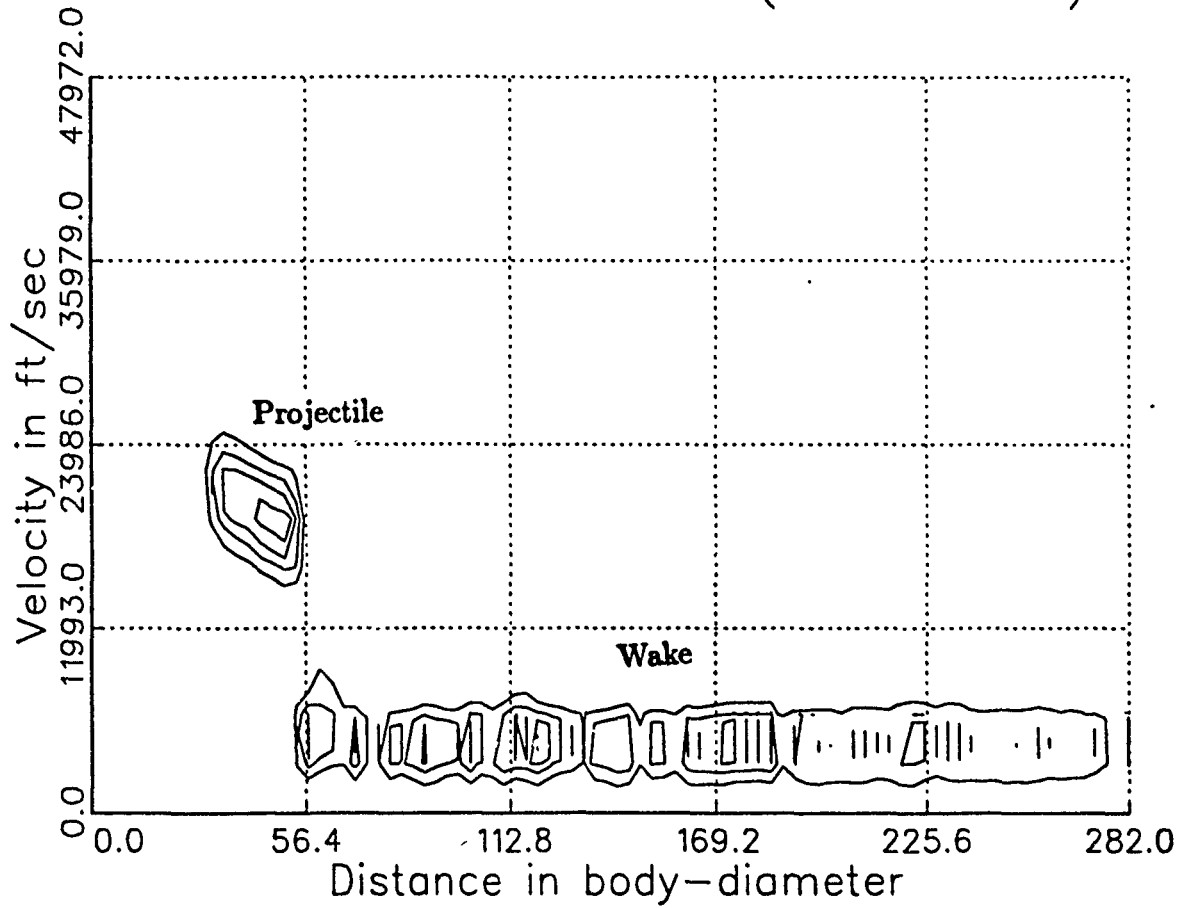


Figure 7: Contour plot of the evolution of doppler frequency with time.

### 3.4 Automatic Extraction of Projectile and Wake

In section 2.4 we discussed the characteristics of the wake and the projectile. In this section we show how these characteristics are used to separate the projectile and the wake from the noise and how the projectile and the wake are automatically separated from each other.

When we observe the raw data we notice that the projectile and the wake have a higher and more constant amplitude than the random noise in the beginning and at the end of the data. Therefore, we locate the maximum point,  $S_{max}$ , in the spectral matrix (Figure 6) and assume it is located on the projectile or the wake (this has always been observed to be the case). Then, we start comparing the maximum peak amplitudes of the  $S_m(k)$  power spectrum estimate from each window to a threshold (for example  $S_{max}/3$  provided good results). The location where the values cross this threshold determines the beginning of the projectile. All the data before that point is considered to be noise.

A similar approach is used to locate the end of the wake. However, we do not compare each peak of the  $S_m(k)$  individually, but rather we create a test window and compare the maximum value of this window to a threshold (for example  $S_{max}/4$  provides good results). Such a window approach is necessary due to the nature of the wake data which may drop below the threshold for a short time. A window will detect whether the data is below the threshold for a longer period of time. Using the two points identified in this way we extract the data from between and consider it to be the data for the projectile plus the wake (See Figure 8).

As was mentioned earlier, after we convert the raw data to velocity we observe a null area immediately behind the projectile. This characteristic plus the fact that the wake should be traveling slower than the projectile provides a way to extract the wake data using a frequency threshold rather than an amplitude threshold. The threshold used here or when the data is being separated from the noise is very dependent on the nature of the data (see Figure 9, the extraction of the wake).

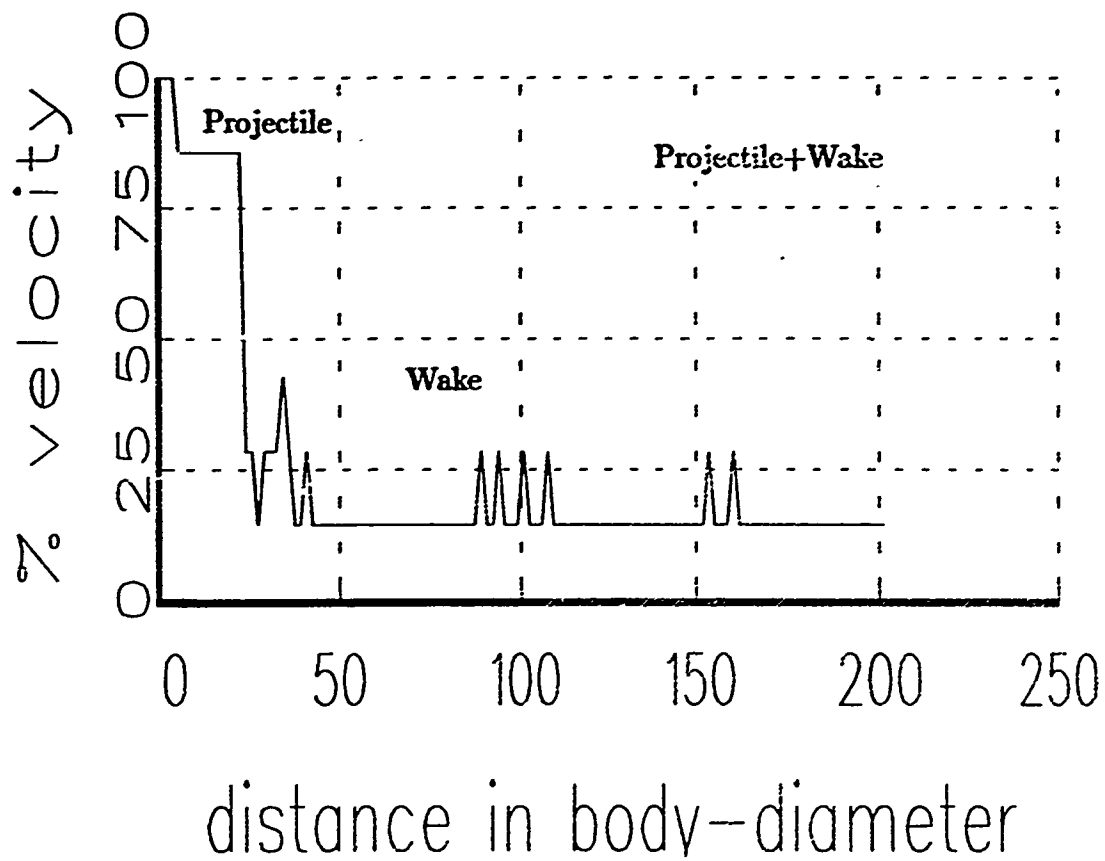


Figure 8: Extracted projectile plus wake velocity.

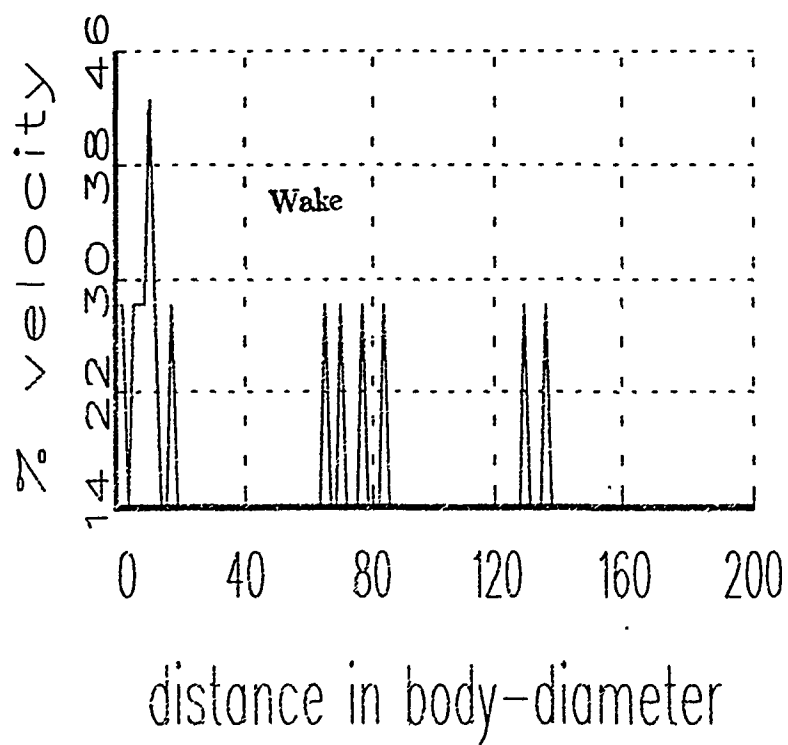


Figure 9: Extracted wake velocity.

## 4 Results

We developed a C program to implement this spectral estimation approach. This program was designed to be user-friendly, and provides an automatic and complete graphical analysis of the velocity behavior.

The user is prompted for a few inputs such as the name of the file containing the measured data, the projectile body diameter, the sampling period, the smoothing window preferred, the radar position and frequency and some other necessary parameters. The C program will then automatically extract the projectile and wake from the data and provide the user with detailed graphical presentations of the velocity behavior of the projectile and the wake. Two of these representations are 2-dimensional plots of velocity versus time. As mentioned earlier the velocities are normalized to the maximum velocity of the projectile and the time is converted to distance in terms of projectile body diameter. The other two graphical representations consist of the 3-dimensional plot of velocity, time, and intensity, and the contour plot. An example of these four plots is given in Figure 10.

These graphical presentations provide the user with revealing views of the data behavior. The program is also capable of comparing the results of a test done using three radars each with a different frequency. The output will be a set of eight plots, four for the projectile plus wake and four for the wake only. This capability allows the comparison of velocity behavior for different frequencies when performed on the same test. This technique can also be used to compare the data for the same frequency but from the In-phase, Quadrature and complex (In-phase and Quadrature data combined) outputs (See Figure 11).

## 5 Conclusion

The proposed approach can be used for automated analysis of doppler radar data. As mentioned in section 4, a relatively small number of initial parameters are needed to provide an automated graphical analysis of the velocity of the projectile and its wake. The proposed means of visualizing the results through 3-D surface plots

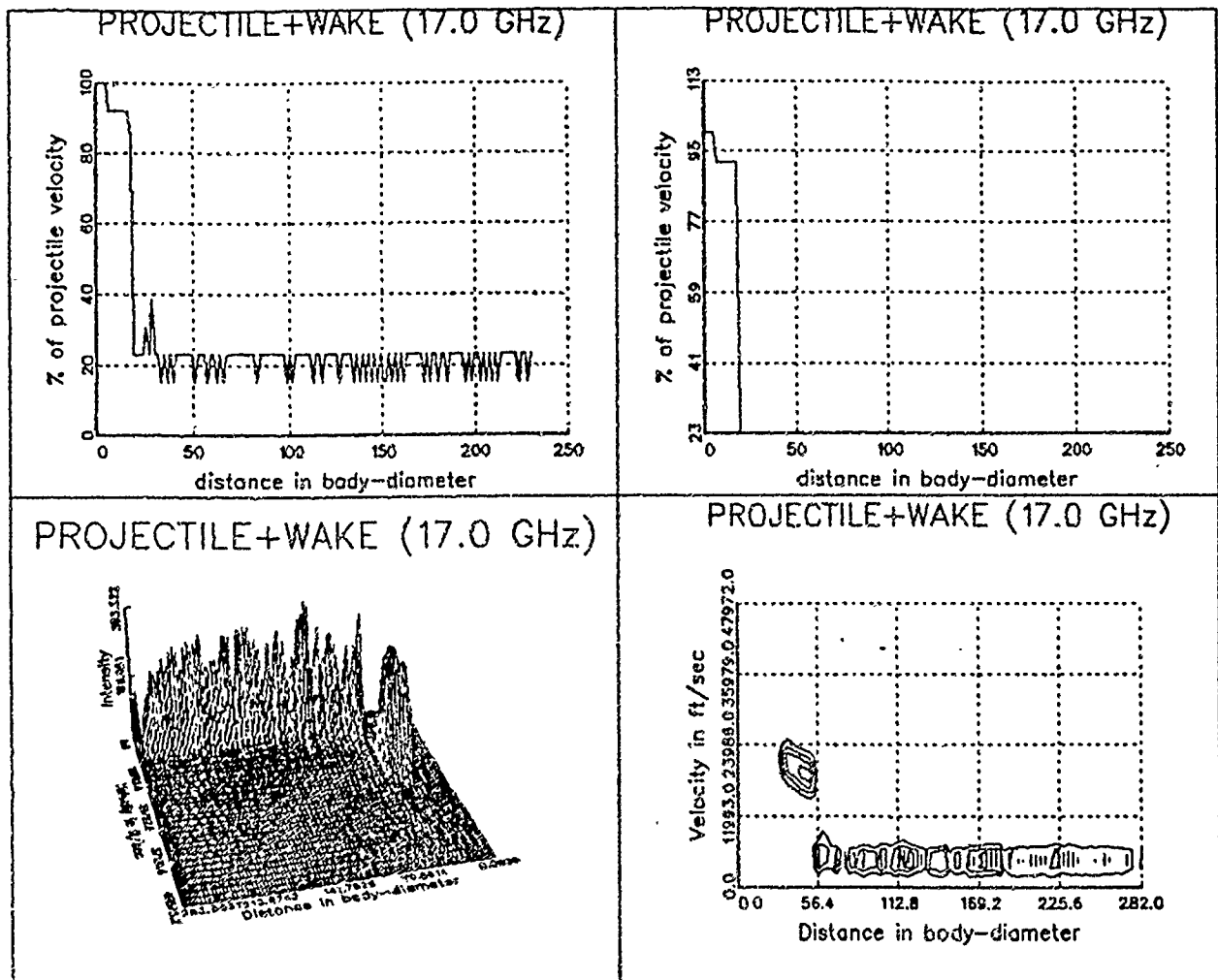


Figure 10: Graphical presentation of results.

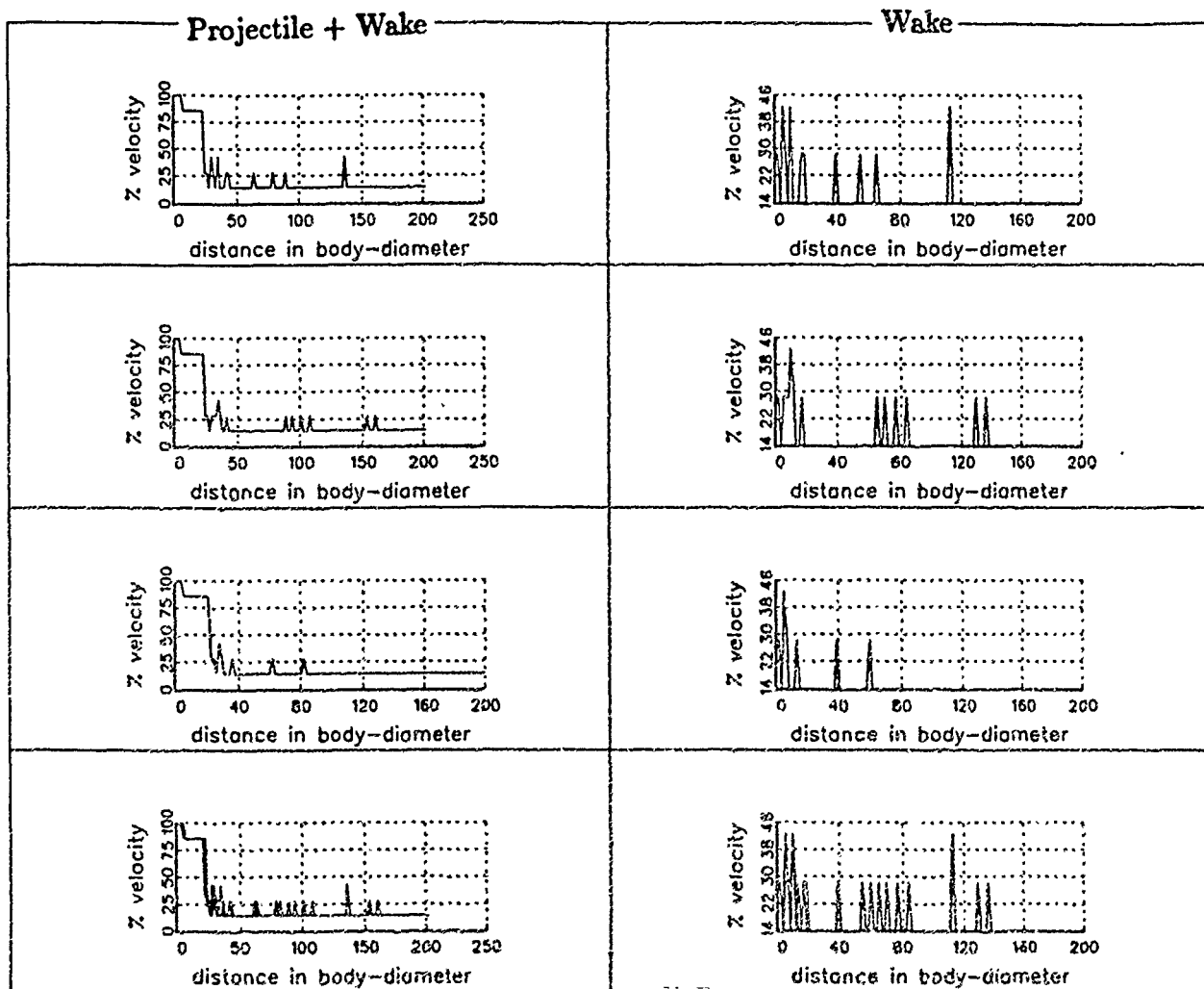


Figure 11: Spectral estimation comparison for 3 sets of data.

and contour plots provide a more complete understanding of the data. Future work may include the use of higher resolution and/or nonstationary spectral analysis techniques such as auto-regressive modeling and time-frequency distribution techniques such as the Wigner-Ville transform that hold promise of increasing the accuracy of the frequency estimates.

Other applications should be studied such as in-barrel doppler data or 3-D imaging of the wake using multiple radars. The latter application is a Phase Interferometer that may be used to characterize qualities of the wake such as its diameter and the density of free electrons.

## References

- [1] *Radar Handbook, 2nd Ed.* ed. by M.I. Skolnik, McGraw-Hill: New York, 1990.
- [2] A.V. Oppenheim and R.W. Schaffer, *Digital Signal Processing*. Prentice-Hall: Englewood Cliffs, 1975.



# AN EXTENDED KALMAN FILTER OBSERVER FOR AN ALTITUDE TEST CELL

D. Mitchell Wilkes and W. Brian Ball  
Vanderbilt University  
Department of Electrical Engineering  
Nashville, TN 37235

### Abstract:

New control techniques are needed to keep pace with the development of technology in altitude test cells. A recently proposed technique for these controls is a model-following controller using an inverse-process model, but there may be room for improvement in the accuracy of the inverse-process model. During this effort, an extended Kalman filter was developed to reduce errors in the inverse-process model.

### Introduction:

As simulated altitude tests are required to become more necessary and economically desirable, the systems required to control these test cells are also having to become more accurate.[1] The technology and techniques now available allow more realistic modelling and estimation of real-world conditions in test cells. Of special interest is the performance testing of engines under transient conditions. With costs of flight testing becoming much more expensive, economical alternatives, especially ground testing, become very important considerations.

The most common controller presently is a PID controller, sometimes gain scheduling is added to the PID controller to increase its flexibility [1]. A recent paper by Chaney [1] proposed an improved technique for controlling tests in these cells. The paper found a model following controller to be the most flexible and accurate form of control in this situation. The control consisted of a second order linear model generating smooth trajectories for each setpoint, using measurements or estimates of process states and parameters in an inverse process model to establish the control inputs to the plant. The controller compensates for nonlinear plant dynamics by linearizing and decoupling the system. Perfect modelling is the goal, but

tracking errors in the system develop from unmodelled dynamics or parameter errors in the feedforward controller. [1]

One source of error results from errors in the observation and estimation of the necessary parameters for the inverse process model. These errors can be reduced by employing an extended Kalman filter observer to estimate these parameters from the observed data. Such a Kalman filter was developed during this effort.

The next section provides a general discussion of the problem at hand. The sections following provide a general procedure for applying an extended Kalman filter to this problem and a specific procedure for developing the Kalman filter is given. A results section then shows the performance of the Kalman filter against simulated data.

#### Problem Discussion:

Our task was to design an extended Kalman filter to observe the important parameters of a altitude test cell. The Kalman filter observer will become part of a controller for the test cell. [1]

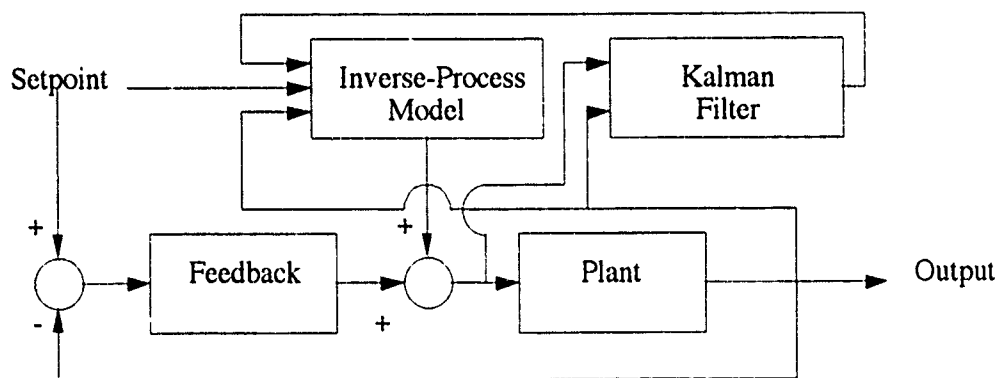


Figure 1 - Model Following Controller with Kalman Filter

The controller shown in Figure 1 is a model-following controller that includes a feedforward inverse-process model to translate the setpoint into

the proper inputs to the test cell. The Kalman filter is necessary to provide high quality estimates of the important parameters so that the feedforward inverse-process model will perform accurately. The test cell under consideration is shown in Figure 2.

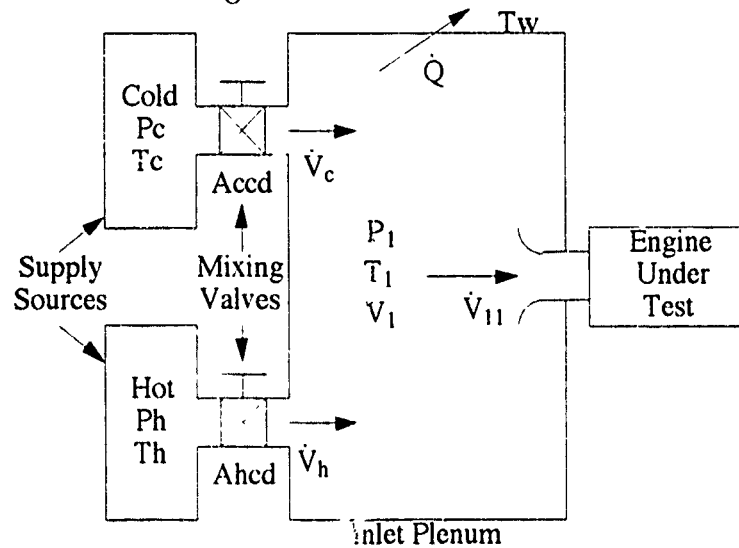


Figure 2: Schematic of Test Cell

The important parameters of the test cell are:

$P_1(t)$	pressure in the inlet plenum
$T_1(t)$	temperature in the inlet plenum
$\dot{V}_{11}(t)$	volume flow rate into the test article
$\dot{Q}(t)$	heat flow rate out of the inlet plenum
$T_w(t)$	temperature of the inlet plenum wall
$\dot{V}_c(t)$	volume flow rate of cold air
$\dot{V}_h(t)$	volume flow rate of hot air
$T_c$	temperature of cold air supply (assumed constant)
$T_h$	temperature of hot air supply (assumed constant)
$A_{ccd}$	valve area ratio for cold air
$A_{hcd}$	valve area ratio for hot air
$\Delta P$	difference between total and static bellmouth pressures
$P_c$	pressure of cold air supply (assumed constant)
$P_h$	pressure of hot air supply (assumed constant)

Table 1 - Parameters

The nonlinear differential equations governing the system are:

$$\dot{P}_1(t) = \frac{\gamma}{V_1} P_1(t) [\dot{V}_h(t) + \dot{V}_c(t) - \dot{V}_{11}(t)] + \frac{\dot{Q}(t)}{V_1} (\gamma - 1)$$

$$T_1(t) = \frac{T_1(t)}{V_1} \left[ \dot{V}_h(t) \left( \gamma - \frac{T_1(t)}{T_h} \right) + \dot{V}_c(t) \left( \gamma - \frac{T_1(t)}{T_c} \right) - \dot{V}_{11} (\gamma - 1) + \frac{\dot{Q}(t)}{P_1(t)} (\gamma - 1) \right]$$

$$\dot{Q}(t) = Ch \left( \frac{P_1(t) \dot{V}_{11}(t)}{T_1(t)^{1.76}} \right)^b As (T_w(t) - T_1(t))$$

$$\dot{V}_{11}(t) = \frac{R T_1(t)}{P_1(t) - \Delta P} \frac{P_1(t) A_{cd}}{\sqrt{T_1(t)}} \sqrt{\frac{2 \gamma}{R (\gamma - 1)}} \sqrt{\left( \frac{P_1(t) - \Delta P}{P_1(t)} \right)^{\frac{2}{\gamma}} - \left( \frac{P_1(t) - \Delta P}{P_1(t)} \right)^{\frac{\gamma + 1}{\gamma}}}$$

$$T_w(t) = \frac{\dot{Q}(t)}{3600 C_m M}$$

$$\dot{V}_h(t) = \frac{R T_h}{P_1(t)} \frac{P_h A_{hcd}}{\sqrt{T_h}} \sqrt{\frac{2 \gamma}{R (\gamma - 1)}} \sqrt{\left( \frac{P_1(t)}{P_h} \right)^{\frac{2}{\gamma}} - \left( \frac{P_1(t)}{P_h} \right)^{\frac{\gamma + 1}{\gamma}}}$$

$$\dot{V}_c(t) = \frac{R T_c}{P_1(t)} \frac{P_c A_{ccd}}{\sqrt{T_c}} \sqrt{\frac{2 \gamma}{R (\gamma - 1)}} \sqrt{\left( \frac{P_1(t)}{P_c} \right)^{\frac{2}{\gamma}} - \left( \frac{P_1(t)}{P_c} \right)^{\frac{\gamma + 1}{\gamma}}}$$

where the meanings of the variables are given in the following table:

$\gamma$	ratio of specific heats
Ch	collection of constants
$A_{cd}$	bell-mouth throat area
b	empirical constant
As	inlet plenum effective surface area
$C_m$	heat capacitance of the metal
R	gas constant
M	mass of the metal

Table 2 - Constants

### Procedure:

Two different formulations of these differential equations were tried in order to strike a good balance of accuracy, utility, and computational complexity. Both formulations required extensive symbolic differentiation in order to derive and linearize the system model before implementing the Kalman filter. This symbolic differentiation was too extensive to handle by hand, therefore a symbolic mathematics package, Mathematica, was employed to compute the derivatives. Each formulation, however, followed the same basic procedure. [2] That basic procedure can be described as follows:

1. From the important parameters of the system model equations, select a set of these parameters (and their time derivatives) to be the state variables of the system. Select the remaining important parameters to be inputs. This results in the nonlinear state equations

$$\dot{\underline{x}}(t) = f(\underline{x}(t), \underline{u}(t))$$

where  $\underline{x}(t)$  is the vector of state variables and  $\underline{u}(t)$  is the vector of input variables. Also the vector of observations became

$$\underline{z}(t) = H^T \underline{x}(t)$$

where the transpose of H selects the observable state variables.

2. The equations for the time derivatives of the state variables define nonlinear differential equations describing the dynamics of the system. Discretize these equations using the Euler approximation to the derivative

$$\dot{\underline{x}}(t) \approx \frac{1}{T_s} [\underline{x}(t+T_s) - \underline{x}(t)]$$

This results in a set of nonlinear difference equations of the form

$$\underline{x}((k+1)T_s) = \underline{x}(k T_s) + T_s f(\underline{x}(k T_s), \underline{u}(k T_s))$$

3. These nonlinear difference equations are then linearized about the state variables and the input variables to obtain a set of linear time-varying difference equations. This will be described in more detail later in this report.
4. The linearized difference equations are used to produce the extended Kalman filter.

The difference in the formulations derives from choosing different sets of state variable , step 1. above.

### Results:

For the first realization, the state vector  $\underline{x}$  was selected as

$$\underline{x}(t) = \begin{bmatrix} P_1(t) \\ T_1(t) \\ \dot{Q}(t) \\ \dot{V}_{11}(t) \\ T_w(t) \\ \dot{V}_h(t) \\ \dot{V}_c(t) \end{bmatrix}$$

and the input vector  $\underline{u}$  was selected as

$$\underline{u}(t) = \begin{bmatrix} P_h(t) \\ P_c(t) \\ T_h(t) \\ T_c(t) \\ A_{hcd}(t) \\ A_{ccd}(t) \\ \Delta P(t) \end{bmatrix}$$

This formulation required that time derivatives of some of the original nonlinear equations be taken. In particular we took time derivatives of the equations for  $\dot{Q}(t)$ ,  $\dot{V}_{11}(t)$ ,  $\dot{T}_w(t)$ ,  $\dot{V}_h(t)$  and  $\dot{V}_c(t)$  to obtain equations for  $\ddot{Q}(t)$ ,  $\ddot{V}_{11}(t)$ ,  $\ddot{T}_w(t)$ ,  $\ddot{V}_h(t)$ , and  $\ddot{V}_c(t)$ . These derivatives resulted in very large and complex

nonlinear differential equations. The resultant vector of nonlinear differential equations,  $\underline{f}$ , is given by

$$\underline{f}(\underline{x}(t), \underline{u}(t)) = \begin{bmatrix} \dot{P}_1(t) \\ \dot{T}_1(t) \\ \ddot{Q}(t) \\ \dot{\tilde{V}}_{11}(t) \\ \dot{\tilde{T}}_w(t) \\ \dot{\tilde{V}}_h(t) \\ \dot{\tilde{V}}_c(t) \end{bmatrix}$$

These nonlinear equations had to be discretized and then linearized to form a set of linear state equations in order to apply the Kalman filter equations. [2] The discretization followed the approach in step 2 above. After dropping the  $f_s$  notation in the arguments of  $\underline{x}(k T_s)$  and  $\underline{u}(k T_s)$  we obtain

$$\underline{x}(k+1) = \underline{x}(k) + T_s \underline{f}(\underline{x}(k), \underline{u}(k))$$

The linearization was carried out by taking derivatives with respect to  $\underline{x}$  to form a state transition matrix  $F$  and with respect to  $\underline{u}$  to form the input coupling matrix  $B$ . The linearized state equations were thus given by

$$\underline{x}(k+1) = F(k) \underline{x}(k) + B(k) \underline{u}(k)$$

where

$$F(k) = \frac{\partial}{\partial \underline{x}(k)} [\underline{x}(k) + T_s \underline{f}(\underline{x}(k), \underline{u}(k))]$$

$$B(k) = \frac{\partial}{\partial \underline{u}(k)} [\underline{x}(k) + T_s \underline{f}(\underline{x}(k), \underline{u}(k))]$$

The measurement equation was given by

$$z(k) = H^T \underline{x}(k) + v(k)$$

where

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$



and  $\underline{v}(k)$  represents uncertainties (i.e. noise and/or error) in the measurements.

The Kalman filter variables, summarized in the table below, were given initial values so that the filter could run.

$S_p$	predictive noise covariance matrix
$S_c$	corrective noise covariance matrix
$H$	output coupling matrix
$R_n$	output noise covariance matrix
$Q_{ns}$	input noise covariance matrix
$\underline{z}$	readings from actual test cell
$\underline{x}_p$	predictive state vector
$\underline{x}_c$	corrective state vector

Table 3 - Kalman Filter Variables

The Kalman filter operated according to the following equations. The Kalman filter first calculated the Kalman gain  $L$  [2]

$$L(k) = S_p(k) H [H^T S_p(k) H + R_n(k)]^{-1}$$

The corrected noise covariance matrix  $S_c$  was then calculated according to [2]

$$S_c(k) = S_p(k) - L(k) H^T S_p(k)$$

The vector of measurements  $\underline{z}$  and the input vector  $\underline{u}$  were formed by reading the following values from the data files (of course, in a real-time environment these would come from the sensors directly).

$$\underline{z} = \begin{bmatrix} P_{lsim}(k) \\ T_{lsim}(k) \\ T_{wsim}(k) \end{bmatrix}$$

$$\underline{u} = \begin{bmatrix} P_h \\ P_c \\ T_h \\ T_c \\ A_h(k) \\ A_c(k) \\ \Delta P(k) \end{bmatrix}$$

The pressures and temperatures of the supply sources  $P_h$ ,  $P_c$ ,  $T_h$ , and  $T_c$  are assumed to be constant.

The corrected state estimate  $\underline{x}_c(k)$  (corrected using the measurements) was given by

$$\underline{x}_c(k) = \underline{x}_p(k) + L(k) [z(k) - H^T \underline{x}_p(k)]$$

The values from  $\underline{x}_c$  were used to evaluate  $F$  and  $B$  at time  $k$  since  $F$  and  $B$  are functions of  $\underline{x}(k)$  and  $\underline{u}(k)$ .  $F$  and  $B$  were then used to calculate the one-step predicted noise covariance matrix  $S_p(k)$ .

$$S_p(k) = F(k) S_c(k) F^T(k) + B(k) Q_{ns} B^T(k)$$

The prediction of the next value of each of the state variables was made from the previously derived discrete nonlinear equations.

$$\underline{x}_p(k+1) = \begin{bmatrix} P1m(k+1) \\ T1m(k+1) \\ \dot{Q}m(k+1) \\ \dot{V}_{11}m(k+1) \\ Twm(k+1) \\ \dot{V}_hm(k+1) \\ \dot{V}_cm(k+1) \end{bmatrix} = \underline{x}_c(k) + T_s f(\underline{x}_c(k), \underline{u}(k))$$

This process repeated over all the points in the simulation data.

The model from above was used with the derivatives of the equations for five of the state variables. The filter worked very well when simulated in the Mathematica package, but the equations for the  $F$  and  $B$  matrices were completely intractable for coding into FORTRAN (the equations describing the elements of the  $F$  and  $B$  matrices required 160 pages to print out), therefore a simpler model was developed. Instead of using  $\dot{Q}$ ,  $\dot{V}_{11}$ ,  $\dot{Tw}$ ,  $\dot{V}_h$  and  $\dot{V}_c$  as the state variables, their integrals  $Q$ ,  $V_{11}$ ,  $Tw$ ,  $V_h$ , and  $V_c$  were used. This simplified the equations greatly, since the need for additional differentiation (to obtain  $\ddot{Q}(t)$ ,  $\ddot{V}_{11}(t)$ ,  $\ddot{Tw}(t)$ ,  $\ddot{V}_h(t)$ , and  $\ddot{V}_c(t)$  in the previous formulation) was eliminated. The model worked well in this simplified

form, but had problems tracking the wall temperature  $T_w$ . This was apparently due to the fact that the diagonal element of  $S_p$  that allows correction from the sample of  $T_w$  was becoming very tiny, causing very little of the correction to affect  $T_w$ . When the engine began a maneuver,  $T_w$  would track poorly because the Kalman gain for  $T_w$  was too tiny to produce much correction. We solved the problem by adding a value of 0.5 to this element of  $S_p$  at every iteration. This resulted in a higher Kalman gain for  $T_w$  thus allowing the sample of  $T_w$  to provide more correction to the value of the predicted  $T_w$  in  $\hat{x}_p(k)$ . See the graphs provided in the appendix.

### Conclusion:

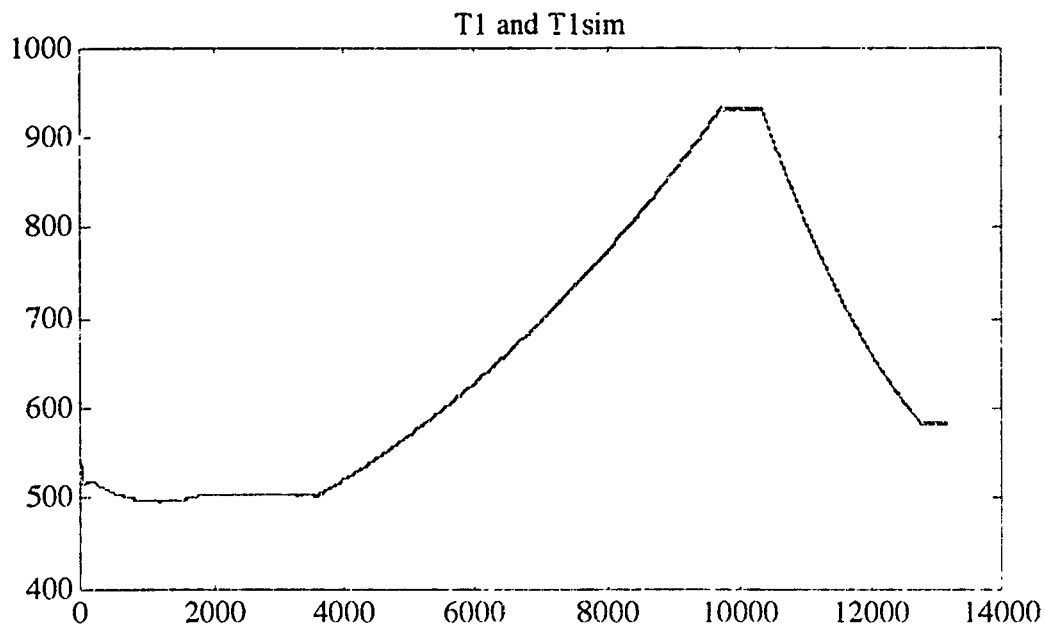
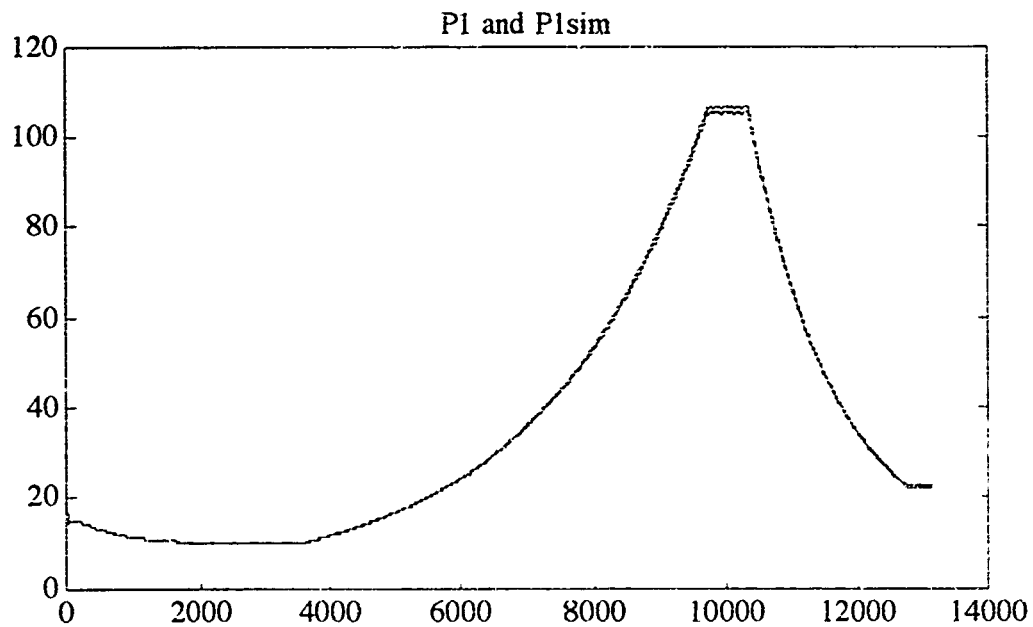
The developed extended Kalman filter tracked very well, but several topics need to be addressed. Future research should include the following topics:

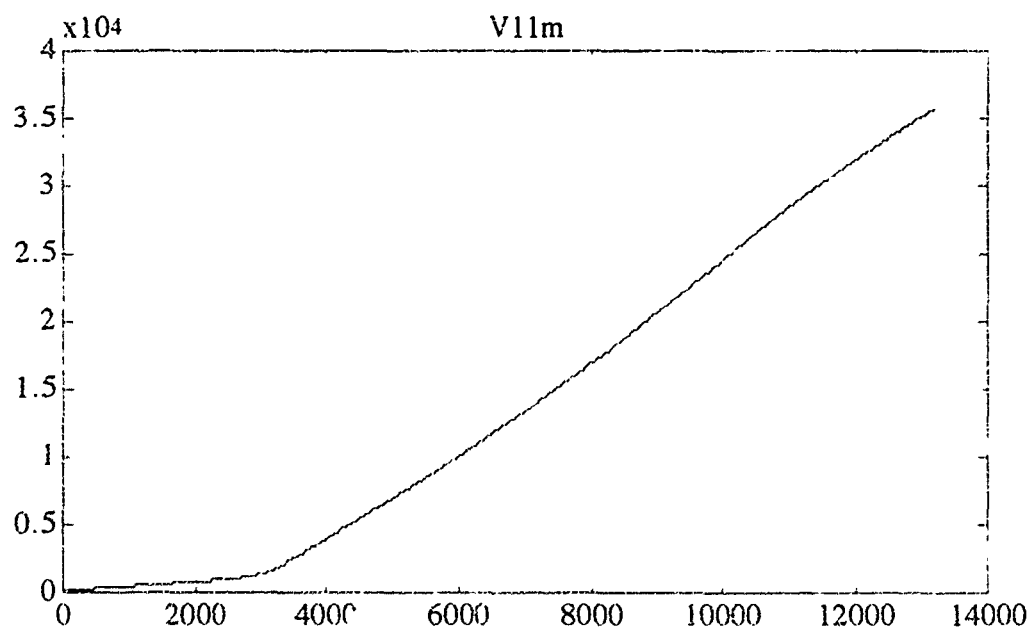
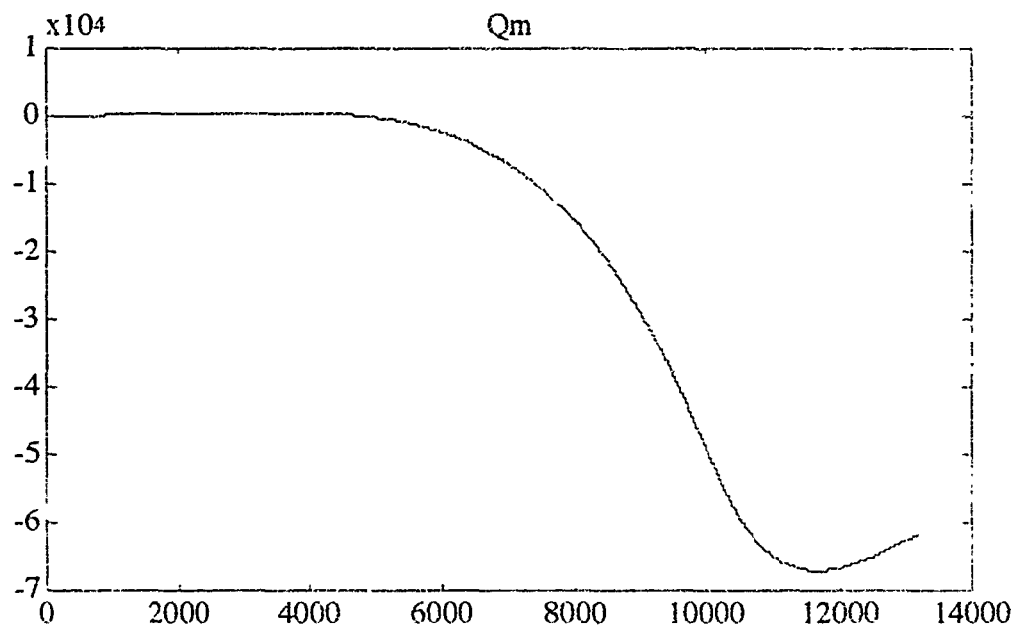
1. Integration of the FORTRAN implementation of the extended Kalman filter into existing simulation and control software to test performance more rigorously and evaluate whether the accuracy level is acceptable.
2. Evaluate whether the present implementation is fast enough for a real-time control situation.
3. Investigate faster implementations using multiprocessors and/or high speed processors.
4. Investigate the stability and sensitivity of the extended Kalman filter to disturbances and modelling error.

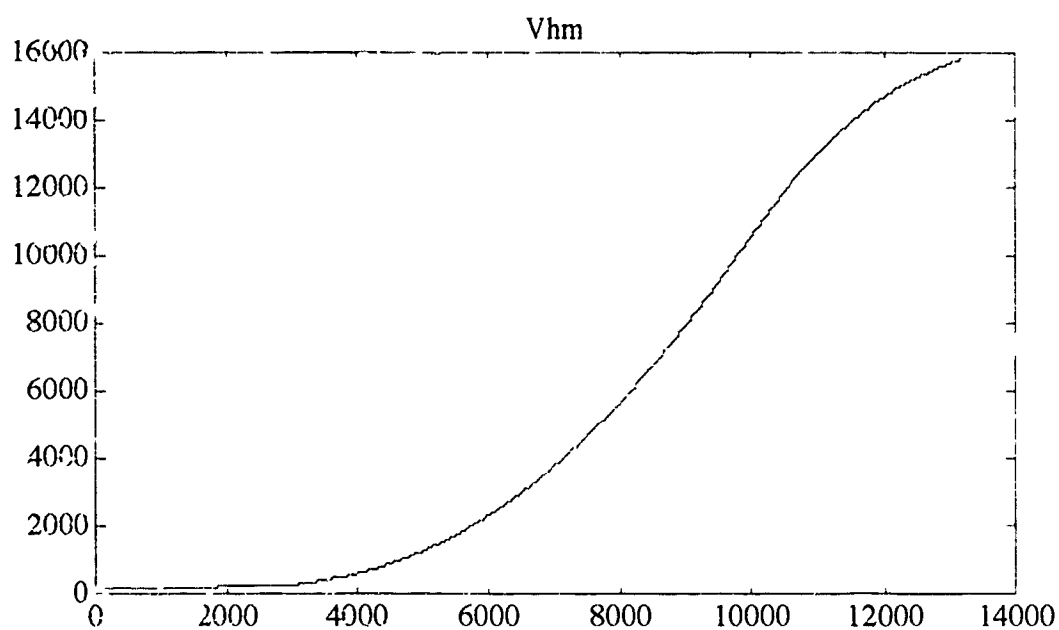
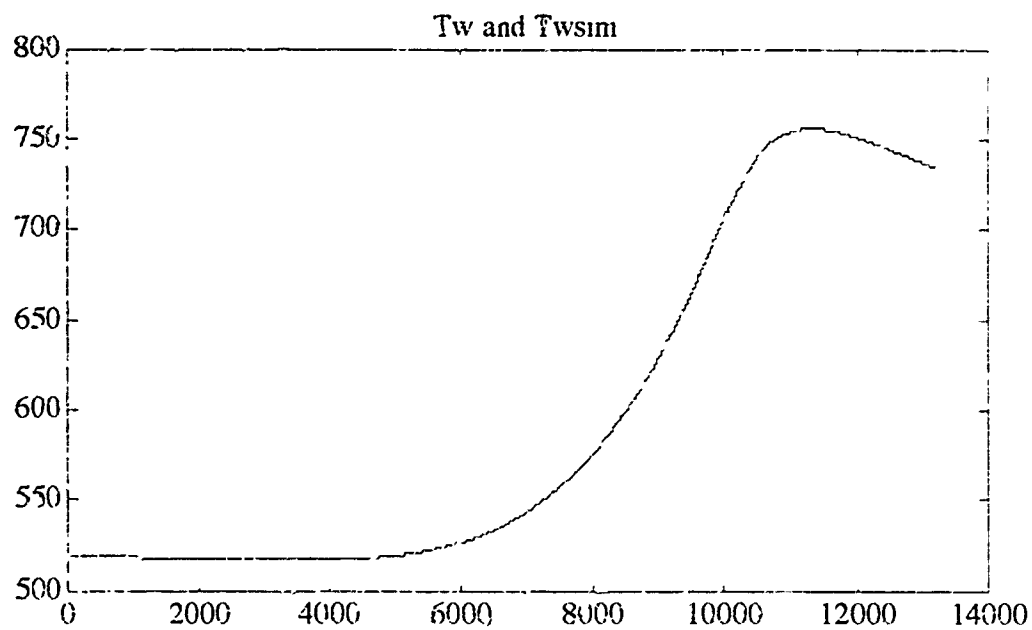
The results of this effort were very encouraging and strongly suggest that the use of an extended Kalman filter can improve the accuracy of the inverse-process model in the model-following controller.

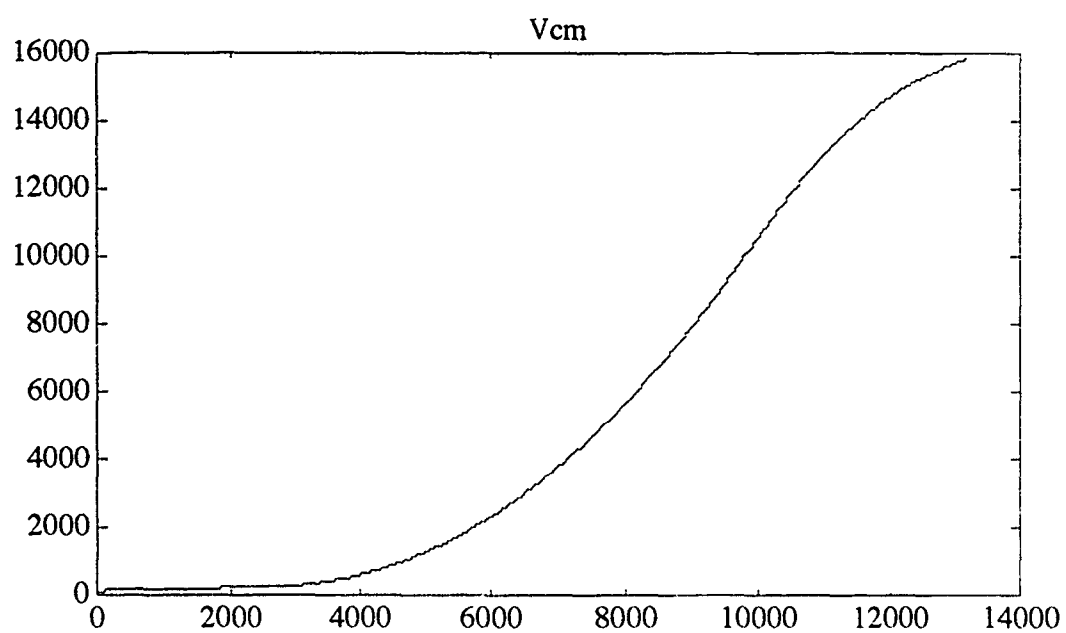
## Appendix

The graphs below show the actual data from P1, T1, and Tw plotted with the predicted values of P1, T1, and Tw. The plots for Q,  $V_{11}$ ,  $V_h$ , and  $V_c$  show the behavior expected of the integral of the quantities of interest.









## References

- [1] M. J. Chaney and J. J. Beaman, "Comparison of Nonlinear Tracking Controllers for a Compressible Flow Process", submitted to *ASME Journal of Dynamic Systems, Measurement and Control*
- [2] B. D. O. Anderson and J. B. Moore, *Linear Optimal Control*, Prentice-Hall, Edgewood Cliffs, N.J., 1971



# NEW REACTION TRANSFORMATIONS USING NITRONIUM TRIFLATE

**Christopher M. Adams, Ph. D.**

Assistant Professor

Department of Chemistry

Oklahoma State University, Stillwater, Oklahoma 74078-0447

**Major Scott A. Shackelford, Ph. D., USAF**

Frank J. Seiler Research Laboratory (AFSC)

FJSRL/NC, United States Air Force Academy

Colorado Springs, Colorado 80840-6528

## ABSTRACT

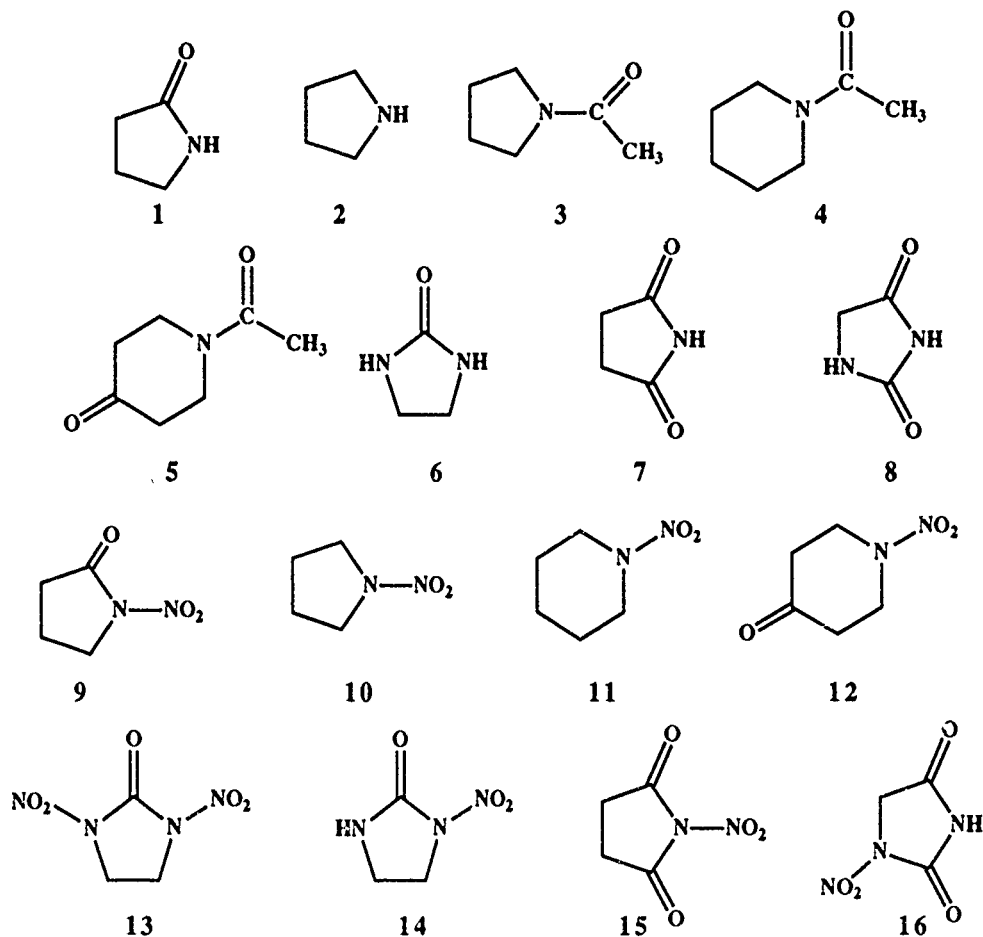
New, non-protic N-nitration methods were explored using nitronium triflate,  $\text{NO}_2\text{OSO}_2\text{CF}_3$  ( $\text{NO}_2\text{OTf}$ ), generated in situ from tetrabutylammonium nitrate,  $\text{Bu}_4\text{NNO}_3$ , and trifluoromethanesulfonic anhydride,  $(\text{CF}_3\text{SO}_2)_2\text{O}$  ( $\text{Tf}_2\text{O}$ ). 2-Pyrrolidinone, pyrrolidine, N-acetyl-pyrrolidine, N-acetyl-piperidine, N-acetyl-4-piperidinone, imidazolidinone, succinimide, and hydantoin were nitrated using these methods. Aqueous and non-aqueous workups gave yields ranging from 20-76%. The reaction methods are significantly improved and eliminate the hazard potential, catalytic and solubilizing necessity and reaction workup requirements of nitromethane as a reaction solvent previously reported. Investigations were performed at the F. J., Seiler Research Laboratory, AFSC FJSRL/NC, U. S. Air Force Academy, Colorado Springs, Colorado.

The nitration of heterocyclic compounds is of great importance to several fields of chemistry. Of primary interest to the armed services is the use of nitro compounds in explosives and propellants.<sup>1-4</sup> The formation of these compounds in large quantities is of prime importance for their production however in the laboratory more convenient methods of nitration are required in order to form trial compounds for study. The use of often harsh conditions exemplifies their general bulk syntheses and many investigators who are interested in the decomposition and stability of the compounds search for milder synthetic conditions for smaller quantities. These conditions would allow modification of the trial compounds and the incorporation of more sensitive functionalities. Thus, there are several areas to be addressed in this study:

- 1) Investigate new alternative methods for the formation of nitronium triflate,  $\text{NO}_2\text{OSO}_2\text{CF}_3$ , ( $\text{NO}_2\text{OTf}$ ), from tetrabutylammonium nitrate,  $\text{Bu}_4\text{NNO}_3$ , and trifluoromethanesulfonic anhydride,  $(\text{CF}_3\text{SO}_2)_2\text{O}$ .
- 2) Investigate non-protic nitration methods.

- 3) Increase the reaction yields and  $\text{NO}_2\text{OTf}$  solubility in the reactions.
- 4) Investigate N-nitration versus C-nitration possibilities.
- 5) Eliminate the catalytic necessity or solubilizing ability of nitromethane.
- 6) Eliminate the hazard potential and reaction workup requirements of nitromethane.
- 7) Determine whether a base is required for intermediate formation in the nitration mechanism.
- 8) Retain  $\text{CH}_2\text{Cl}_2$  as a reaction solvent for other reactions and then substitute alkane solvents for reactions which cannot be run in chlorinated solvents.

The heterocyclic compounds (1 - 8) and their corresponding nitro derivatives (9 - 16) used or synthesized in this study are given below:



## Nitronium Source

The introduction of nitro groups into organic molecules has been done using a variety of methods.<sup>5-7</sup> Most methods require strongly acidic conditions, normally using oxy-acids. The strongly acidic conditions react with the often sensitive functionalities, limiting application of the reactions in specific compound syntheses. These conditions also induce polymerization making the method not applicable to polymer precursors for energetic binders. Recently, the use of trifluoroacetic anhydride and ammonium nitrate in pure nitromethane as a solvent has been reported by Suri and Chapman.<sup>5</sup> These conditions advanced the use of more convenient methodology for nitration but suffered from the requirement of nitromethane as a solvent. While nitromethane is required to solubilize the starting reagents and the subsequent nitronium salts, Major Scott A. Shackelford, at the F. J. Seiler Research Laboratory at the U.S. Air Force Academy, postulated the use of tetrabutyl ammonium nitrate as the nitrate source due to its increased solubility in methylene chloride and the use of trifluoromethanesulfonic anhydride as the nitrate deoxygenating agent due to the better leaving ability of triflate anion.

There are several advantages in the use of tetrabutylammonium nitrate. The quaternary ammonium salt is not a proton source. The quaternary ammonium cation can act as a phase transfer agent solubilizing nitrate and remove the requirement for nitromethane in known nitration reactions. And, the workup may be easier as the ammonium salt can be removed by chromatography.

## Heterocycle Investigations

The structures of the heterocycles to be used in this study vary considerably in the chemical availability of the nitrogens for nitration. On examination there is a secondary amine (2), four amides (1, 3, 4, 5), one imide (7), one imidazole (6) and one oxazole (8). The placement and affect of the carbonyl groups can contribute to the availability of the free electron pair on the nitrogen. How do these structure variations affect the reactivity and yield of the nitration reactions? Do these variations show up in different isolation techniques? Can these these compounds act as a base in the reactions, i.e. neutralize the reaction using the substrate, and does the addition of a weak base assist in the nitration reactions?

## 2-Pyrrolidinone (1) Investigations: Formation of 1-Nitro-pyrrolidinone (9)

Table 1. Synthesis Reactions Producing 1-Nitro-Pyrrolidinone (9).

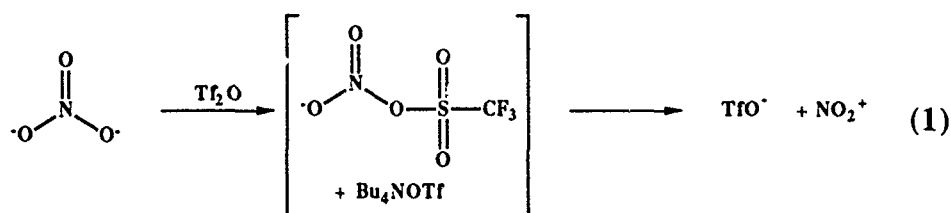


Entry	Solvent of Generation	T (°C)	Salt R of $\text{R}_4\text{NNO}_3$	Ratio Rgt/Subs	Solvent of Rxn	°C, time	Yield	Run
1	$\text{CH}_2\text{Cl}_2$	-20	H	1:1	$\text{CH}_2\text{Cl}_2/\text{Na}_2\text{SO}_4$	-20, 1 hr	12	1A1-1
2						-20, 1 hr	13	1B1-9
3	$\text{CH}_2\text{Cl}_2$	-20	H	2:1	$\text{CH}_2\text{Cl}_2/\text{Na}_2\text{SO}_4$	-20, 1 hr	16	2A1-3
4	$\text{CH}_3\text{NO}_2$	-20	H	1:1	$\text{CH}_2\text{Cl}_2/\text{Na}_2\text{SO}_4$	-20, 1 hr	31	1D1-32
5	$\text{CH}_2\text{Cl}_2$	-20	Bu	1:1	$\text{CH}_2\text{Cl}_2$	-20, 1 hr	43	6A1-20
6	$\text{CH}_2\text{Cl}_2$	-20	Bu	1:1	$\text{CH}_2\text{Cl}_2/\text{Na}_2\text{SO}_4$	-20, 1 hr	38	8A1-25
7	$\text{CH}_2\text{Cl}_2$	-20	Bu	2:1	$\text{CH}_2\text{Cl}_2$	-20, 1 hr	41	7A1-23
8	$\text{CH}_3\text{NO}_2$	-20	Bu	1:1	$\text{CH}_2\text{Cl}_2$	-20, 1 hr	45	11A1-30
9	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	0-RT, 17 hr	56.5	16A1-42
10	$\text{CH}_2\text{Cl}_2$	0	Bu	1:2	$\text{CH}_2\text{Cl}_2$	RT, 17 hr	63(Tf)*	20A2-12
11	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	RT, 1 hr	31.5(Pyr)* 25.3	24A2-20
12	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	RT, 1 hr, Naq.	0	27A2-38

\* Yield of 63% based on nitration reagent. Yield of 31.5% based on 2-pyrrolidinone.

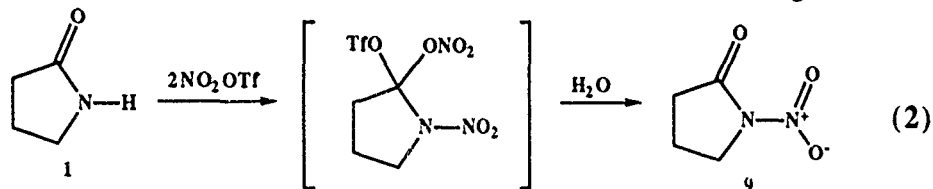
Pyrrolidinone (1) has been used as a standard for amide nitrations. Literature reports on pyrrolidinone nitrations using a variety of reagents range from 30 to 56%.<sup>6</sup> Using trifluoroacetic anhydride and ammonium nitrate Suri and Chapman report a yield of 30% for 9.

To establish a standard ammonium nitrate,  $\text{NH}_4\text{NO}_3$ , and tetrabutylammonium nitrate,  $\text{Bu}_4\text{NNO}_3$ , were allowed to react with trifluoromethanesulfonic anhydride, triflic anhydride ( $\text{Tf}_2\text{O}$ ), in pure methylene chloride to generate nitronium triflate, equation 1. As expected the insolubility of the  $\text{NH}_4\text{NO}_3$  in methylene chloride gave low yields, 13%, Table 1. However, the more soluble  $\text{Bu}_4\text{NNO}_3$  gave a significant increase to 43%. Use of  $\text{Na}_2\text{SO}_4$  as a weak base to neutralize any acid produced when  $\text{NH}_4\text{NO}_3$  is the nitrate source showed no significant increase in yield, 16%. When  $\text{Na}_2\text{SO}_4$  was used in the reaction with  $\text{Bu}_4\text{NNO}_3$  as the nitrate source the yield was comparable to that without  $\text{Na}_2\text{SO}_4$ , 38%. This would suggest that the protons derived from the ammonium cation or the amide do not interfere with the nitration reaction.



To determine whether nitromethane serves as a catalyst and/or just a solubilizing agent, each reagent was examined using 8 equivalents of nitromethane. The use of 8 equivalents was experimentally determined as this was the minimum amount of nitromethane required to stir the reaction mixture during formation of the nitronium triflate. When the substrate was added, methylene chloride was used to help solubilize the reaction mixture. Using  $\text{NH}_4\text{NO}_3$ ,  $\text{Tf}_2\text{O}$ , and  $\text{Na}_2\text{SO}_4$  the yield was 31% which is comparable to using  $(\text{CF}_3\text{CO})_2\text{O}$  and  $\text{NH}_4\text{NO}_3$  in pure nitromethane solvent. It would appear that nitromethane serves a solubilizing function. This also compares favorably with the initial reaction run by Major Shackelford at FJSRL. When  $\text{Bu}_4\text{NNO}_3$  was substituted as the nitrate source, using 8 equivalents of nitromethane to initially generate the nitronium triflate, the yield of **9** was increased to 45%. This can represent a solubilizing contribution by nitromethane on the nitronium triflate salt generated prior to nitration.

It has been suggested by Suri that the substrate may participate as a base in the reaction mixture.<sup>5</sup> If one equivalent of reagent is taken up by the acid produced after nitration then the overall yield would be reduced. Major Shackelford postulated that a twofold addition of nitronium ion to pyrrolidinone may occur, equation 2. Thus a comparison was done using  $\text{Bu}_4\text{NNO}_3$  and doubling the initial substrate concentration. Conversely, a comparison was made doubling the reagent concentration. When the substrate is doubled, and  $\text{NO}_2\text{Tf}$  is the limiting reagent, the substrate can participate more effectively as a base. On comparison of entry 9 (56.5%), and 10 (63%), only a slight increase in yield is observed, insufficient for participation of the substrate as a base. When the reagent,  $\text{NO}_2\text{Tf}$ , is doubled, compare entry 6 (38%), and 7 (41%), little difference is observed. This would suggest that participation of a second  $\text{NO}_2\text{Tf}$  in the reaction does not occur. When one compares the different nitrate salts, the solubility of the salts when the reagent is doubled exhibits more of an effect than the concentration change.

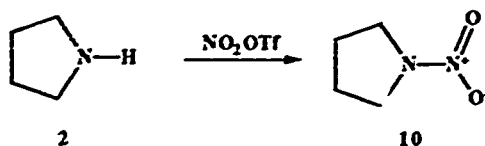


All reactions, irrespective of their incubation temperatures, were allowed to react for 1 hour to form the nitronium triflate reagent prior to addition of the substrate. A comparison of incubation times and yields following substrate addition would thus show a function of the amount of nitronium triflate formed during the 1 hour period. As a matter of note, the nitronium triflate appeared to be a white insoluble salt in all reaction mixtures, as evidenced by the formation of what appeared to be a white solid and a thickening of the reaction mixture. When identical reactions were allowed to proceed at room temperature (RT) for 1 hour and 18 hours following substrate addition, entry 11 (25%) and 9 (57%) showed that longer incubation times would significantly increase the reaction yield. This may be a consequence of solid nitronium triflate being consumed and redissolved as the reaction is allowed to proceed.

The use of a non-aqueous workup in the nitration of 2-pyrrolidinone did not give the desired N-nitro-2-pyrrolidinone. This result is in total disagreement with the reported isolation by Suri and Chapman, but agrees with the observations initially observed by Major Shackelford and Dr. Clay Sharts at FJSRL.<sup>8</sup> Following concentration and chromatography a pale yellow impure solid was obtained. This solid was insoluble in  $\text{CDCl}_3$ , however an NMR in DMSO showed a complex mixture. No starting material or product signals were observed. Signals at  $\delta$  11.97 and  $\delta$  6.08 may indicate an acid and/or an N-H functionality and peaks at  $\delta$  4.37, 2.55, and 2.30 appear close to values for butyrolactone.<sup>9</sup> However, an open chain acid or amine cannot be eliminated, see equation 3.

## Pyrrolidine (2) Investigations: Formation of 1-Nitro-pyrrolidine (10)

Table 2. Synthesis Reactions Producing 1-Nitro-pyrrolidine (10)

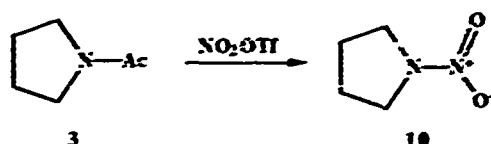


Entry	Solvent of Generation	T (°C)	Salt R of $\text{R}_4\text{N}^+\text{NO}_3^-$	Ratio Rgt/Subs	Solvent of Rxn	°C, time	Yield	Run
1	$\text{CH}_2\text{Cl}_2$	-70	H	1:1	$\text{CH}_2\text{Cl}_2$	-20, 1 hr	51.4	3A1-5
2	$\text{CH}_2\text{Cl}_2$	-20	Bu	1:1	$\text{CH}_2\text{Cl}_2$	-20, 1 hr	43	9A1-27
3	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	RT, 1 hr	38.8	25A2-28
4	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	Rt, 17 hr	33	25B2-30

Pyrrolidine (2) is a secondary amine without the carbonyl associated with the nitrogen for conjugation. Nitronium triflate derived from  $\text{Bu}_4\text{NNO}_3$  and triflic anhydride was used to explore the nitrations of this heterocycle, Table 2. When  $\text{NH}_4\text{NO}_3$  and  $\text{Bu}_4\text{NNO}_3$  were allowed to react, compare entry 1 (51%) and 2 (43%), the results suggest that the ammonium nitrate may be better than the tetrabutyl ammonium salt. This result may be low considering the other reactions explored. When the reaction temperatures are increased there is no significant increase in 1-nitropyrrolidine formation. When the reaction time is increased there may be a slight lowering of the reaction yield but this may be within error. A question of whether the carbonyl contributes to the reactivity of the nitronium triflate as compared to 2-pyrrolidinone can be seen. The carbonyl increases the yield of the nitrations. An anomaly seems to be with temperature of the nitronium triflate formation.

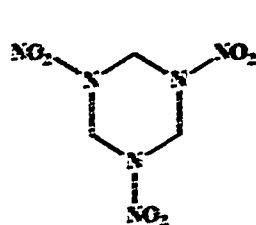
### 1-Acetyl-pyrrolidine (3): Formation of 1-Nitro-pyrrolidine (10)

Table 3. Synthesis Reactions Producing 1-Nitro-pyrrolidine (10)

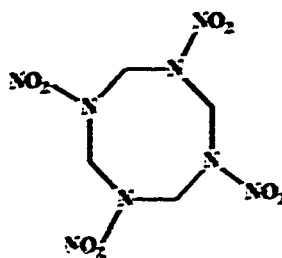


Entry	Solvent of Generation	T (°C)	Salt R of $\text{R}_4\text{NNO}_2$	Ratio Rgt/Subs	Solvent of Rxn	°C, time	Yield	Run
1	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	RT, 1 hr	47.8	29A2-40
2	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	Rt, 18 hr	47.9	29B2-42

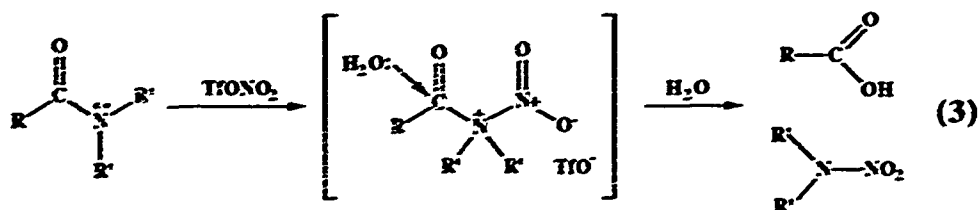
Most nitrations used in the formation of HMX or RDX occur by displacement of an acetylated nitrogen, an amide. In 1-acetyl-pyrrolidine (3) the amide and carbonyl are substituted in a way as to be eliminated upon conversion to nitrated product, Table 3. Hence, if the amide bond in 2-pyrrolidinone is broken an acid may be formed, equation 3. The product may subsequently be eliminated in the workup or by chromatography. Comparison of pyrrolidine with N-acetylpyrrolidine suggests that the carbonyl, in the form of an acetyl, increases the yield of the nitration. This may be a consequence of the intermediates and subsequent breakdown products of the reaction or geometric restrictions enforced by cyclized starting material. When such products are intramolecularly bound where R is attached to R' in equation 3, then the nitration products could be removed in the workup.



**RDX**

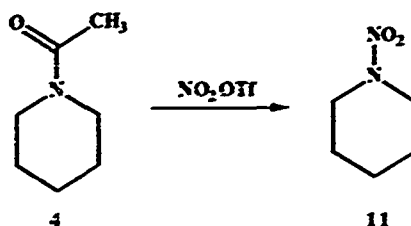


**HMX**



#### 1-Acetyl-piperidine (4): Formation of 1-Nitro-piperidine (11)

Table 4. Synthesis Reactions Producing 1-Nitro-piperidine (11).



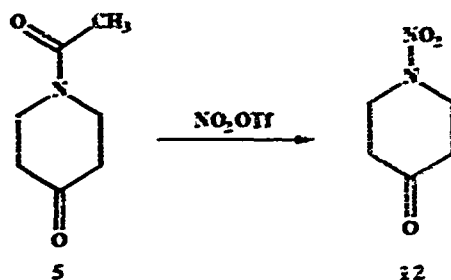
Entry	Solvent of Generation	T (°C)	Salt R of $P_4NNO_3$	Ratio Rgt/Subs	Solvent of Rxn	°C, time	Yield	Run
1	$CH_2Cl_2$	0	Bu	1:1	$CH_2Cl_2$	RT, 1 hr	69	23A2-16
2	$CH_2Cl_2$	0	Bu	1:1	$CH_2Cl_2$	RT, 16 hr	72.3	22B2-22
3	$CH_2Cl_2$	0	Bu	1:1	$CH_2Cl_2$	RT, 1 hr, Naq	53.6	33A3-10

The formation of 1-nitro-piperidine (11) from 1-acetyl-piperidine (4) parallels that of 1-nitro-pyrrolidone, Table 4. The major difference is the size of the ring in the acetylated starting material. The nitrations run from 1 hour and 16 hours at room temperature using an aqueous workup gave yields of 69% and 72%, respectively. This would suggest that the acetyl displacement is nearly complete after 1 hour. If the isolation is performed using non-aqueous conditions the yield is slightly reduced, 64%. In general, the six membered ring starting material gives a 50% increase in isolated yield when compared to 1-acetyl-pyrrolidone.



## 1-Acetyl-4-piperidone (5): Formation of 1-Nitro-4-piperidone (12)

Table 5. Synthesis Reactions Producing 1-Nitro-4-piperidone (12).



Entry	Solvent of Generation	T(°C)	Salt R of $\text{R}_4\text{NNO}_2$	Ratio Rgt/Subs of Rxs	Solvent of Rxn	°C, time	Yield	Run
1	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	RT, 6 hr	22	23A2-18
2	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	RT, 16 hr	18.5	23B2-24
3	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$ , RT, 1 hr, Naq		31.5	32A3-8

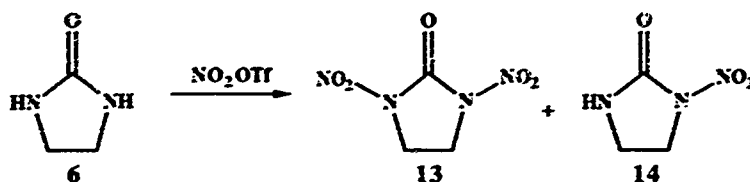
The structure of 1-acetyl-4-piperidone (5) is unique in comparison to the other heterocycles used in this study, Table 5. In all other cases, a carbonyl is adjacent to the nitrogen functionality, i.e. amide functionalities, while in 1-acetyl-4-piperidone (5) a ketone carbonyl is isolated from the nitrogen in addition to the amide. Thus 5 could exhibit through-space electronic contributions to the substitution reactions. Using this analogy 1-acetyl-4-piperidone (5) would be expected to have yields comparable to 1-acetyl piperidine. This is not the case. The yields of 22% (entry 1) and 18% (entry 2) are considerably lower using an aqueous workup and is 31% using a non-aqueous workup. Speculation on the mechanism may suggest that the ring carbonyl in 5 participates in some fashion to the stability of the reaction intermediates and may play a part in intermediate decomposition during an aqueous workup.

## 1-Imidazolidone (6): Formation of 1,3-dinitro-2-oxotetrahydroimidazole (13)

The cyclic urea, 1-imidazolidone (6), has a structure which can allow resonance of a carbonyl with two nitrogen functionalities, Table 6. With two possible reaction sites, two products are possible, 1,3-dinitro-2-oxotetrahydroimidazole (13) and 1-nitro-2-oxotetrahydro-imidazole (14). In the reactions performed, the observed product was 13. Using an aqueous workup yields of 5.6% for 13 was observed for a reaction run entirely at  $-20^\circ\text{C}$ . However, when the temperature was raised to  $0^\circ\text{C}$  for 18 hours, the yield jumped to 19%. Substitution of nitromethane as the reaction solvent gave a comparable

low yield of 5%. This would suggest that reagent solubility with nitromethane was not a factor in the observed yield. When the reaction was run at room temperature for 1 hour and a non-aqueous workup performed, the observed yield of 32.6% was considerably higher than in earlier experiments. Since no mono-nitro product **14** is observed, **14** is more reactive than the starting material. A plausible explanation for the increased reactivity of **14** towards nitration is the formation of a product where the lone pair of the nitrogen in the nitro-amine is in resonance with the carbonyl and the nitro group. This structure would isolate the resonance contributions of the amide linkages and give the unsubstituted amide more amine character, i.e. free the electron pair on nitrogen for substitution.

Table 6. Synthesis Reactions Producing 1,3-dinitro-2-oxotetrahydroimidazole (**13**).



Entry	Solvent of Generation	T(°C)	Salt R of R <sub>1</sub> NNO <sub>2</sub>	Ratio Rgt/Subs	Solvent of Rxn	°C, time	Yield	Run
1	CH <sub>2</sub> Cl <sub>2</sub>	-20	Bu	2:1	CH <sub>2</sub> Cl <sub>2</sub>	-20, 1 hr	5.6	13A1-36
2	CH <sub>2</sub> Cl <sub>2</sub>	0	Bu	2:1	CH <sub>2</sub> Cl <sub>2</sub>	0, 24 hr	19	14A1-38
3	CH <sub>3</sub> NO <sub>2</sub>	0	Bu	2:1	CH <sub>2</sub> Cl <sub>2</sub>	0, 24 hr	5	15A1-40
4	CH <sub>2</sub> Cl <sub>2</sub>	0	Bu	2:1	CH <sub>2</sub> Cl <sub>2</sub>	RT, 1 hr, Naq	32.6	31A3-6

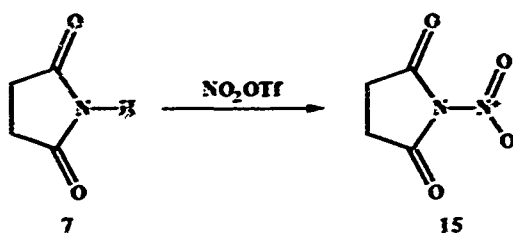
### Succinimide (7): Formation of N-Nitrosuccinimide (15)

In succinimide (**7**) the electron pair in the nitrogen is doubly stabilized by the flanking carbonyl in the amide linkages, Table 7. The acidity of the protons on the atoms adjacent to two carbonyl is well known. Thus the electrons on nitrogen in succinimide would not be expected to be available as in other amides for nitration. When nitration was attempted using an aqueous workup, no N-nitro-succinimide was isolated. On careful inspection during the reaction it appeared that any product formed was reacting further during isolation. When the workup was changed to non-aqueous conditions, the yield increased dramatically to 69% and 76%. Since the only change in reaction conditions was a 5% NaHCO<sub>3</sub> quench and extraction, the result suggest that water destroys the product. In a non-aqueous workup reported by Suri the yield was 28% and clearly our method is significantly better.

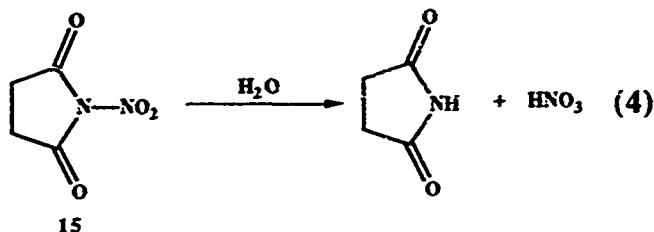
To rationalize these observations one must consider the possible hydrolysis of the N-nitrosuccinimide.<sup>10</sup> As in N-bromosuccinimide, where exposure to HBr produces Br<sub>2</sub> and succinimide, water may be significantly acidic as to react with the product to produce

succinimide and nitric acid, equation 4. This sensitivity and decomposition in water is eliminated by the non-aqueous workup.

**Table 7. Synthesis Reactions Producing N-Nitrosuccinimide (15).**



Entry	Solvent of Generation	Temp (°C)	Salt R of $R_4NNO_3$	Ratio Rgt/Subs	Solvent of Rxn	°C, time	Yield	Run
1	$CH_2Cl_2$	-20	Bu	1:1	$CH_2Cl_2$	-20, 1 hr	0	10A1-29
2	$CH_3NO_2$	9	Bu	1:1	$CH_3NO_2$	RT, 17 hr	?	21A1-14
3	$CH_2Cl_2$	0	Bu	1:1	$CH_2Cl_2$	RT, 1 hr, Naq	69	28A2-36
4	$CH_2Cl_2$	0	Bu	1:1	$CH_2Cl_2$	RT, 18 hr, Naq	75.6	28B2-44

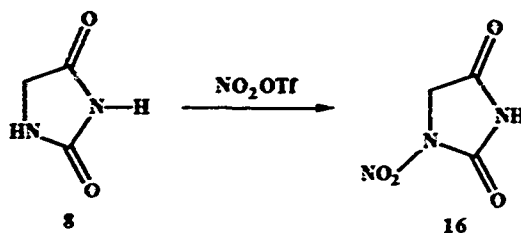


### Hydantoin (8): Formation of 1-Nitrohydantoin (16)

The structure of hydantoin incorporates the features of succinimide and pyrrolidinone, Table 8. Hydantoin has two nitration sites and there was a big difference in workup procedures. While use of nitromethane or methylene chloride did not give substantial amounts of 1-nitrohydantoin in the aqueous workup, a trace amount of product (<5%) was observed when nitromethane was used. Realizing the sensitivity to workup and the observed increased yield with succinimide, hydantoin was thought to have the same characteristics. The reaction of hydantoin, with a non-aqueous workup, gave 69% yield of 16 when no product had been previously isolated. Interestingly, the initial solid isolated following chromatography had a pale yellow color. Storage overnight in a sealed flask resulted in the formation of a brown-orange gas. This would indicate an outgassing of the sample to produce  $NO_2$ . Characterization of the initial isolate was not done, however the changes would suggest that the colored gas is generated from an unstable dinitro compound. The actual structure and mechanism of this decomposition should be explored further. In light of the structures of N-nitrosuccinimide and 1-nitro-pyrrolidinone, the

isolation of mono-nitrated compound with the nitro group on the least reactive amide is expected. However the non-aqueous workup may limit the yield considering the non-aqueous workup of pyrrolidinone.

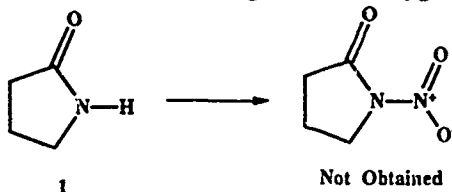
**Table 8. Synthesis Reactions Producing 1-Nitrohydantoin (16).**



Entry	Solvent of Generation	T(°C)	Salt R of $\text{R}_4\text{NNO}_3$	Ratio Rgt/Sub	Solvent of Rxn	°C, time	Yield	Run
1	$\text{CH}_2\text{Cl}_2$	-20	Bu	1:1	$\text{CH}_2\text{Cl}_2$	-20, 1 hr	0	12A1-34
2	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	RT, 17 hr		12B1-44
3	$\text{CH}_3\text{NO}_2$	0	Bu	1:1	$\text{CH}_3\text{NO}_2$	RT, 17 hr		18A2-8
4	$\text{CH}_3\text{NO}_2$	0	H	1:1	$\text{CH}_3\text{NO}_2$	RT, 17 hr	<5	19A2-10
5	$\text{CH}_2\text{Cl}_2$	0	Bu	1:1	$\text{CH}_2\text{Cl}_2$	RT, 1 hr, Naq	69.2	30A3-4

### Explorations of New Nitration Possibilities.

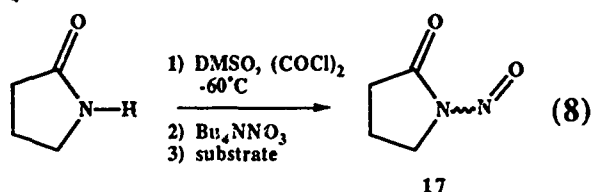
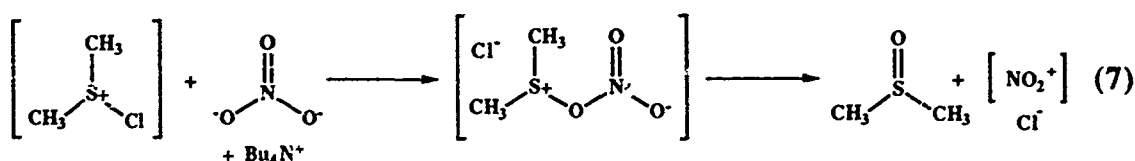
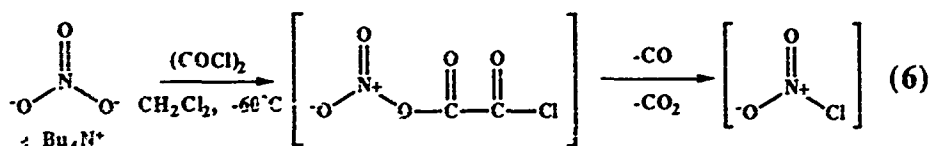
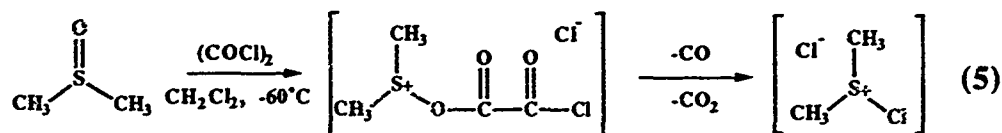
**Table 9. Synthesis Attempts Using New Deoxygenation Techniques.**



Entry	Solvent of Generation	T(°C)	Salt R of $\text{R}_4\text{NNO}_3$	Ratio Rgt/Subs	Solvent of Rxn	°C, time	Yield	Run
1	$\text{CH}_2\text{Cl}_2$	-78	H/SOCl <sub>2</sub>	1:1	$\text{CH}_2\text{Cl}_2$	-78, 1 hr	4	5A1-14
2	$\text{CH}_2\text{Cl}_2$	-60	H/SOCl <sub>2</sub>	1:1	$\text{CH}_2\text{Cl}_2$	-60, -10, 1 hr		5B1-18
3	$\text{CH}_2\text{Cl}_2$	-60	DMSO/(COCl) <sub>2</sub>	1:1	$\text{CH}_2\text{Cl}_2$	-60-RT, 1 hr	0 (no prod)	17A1-46

An examination of the mechanism of nitronium triflate formation gives an overall deoxygenation of nitrate anion, cf equation 1. Similiar reactions have been used to activate carbon dioxide.<sup>11</sup> From this observation it was postulated that other deoxygenation conditions, especially non-protic conditions, may allow deoxygenation of nitrate anion, Table 9. In the Swern oxidation, an active acid chloride deoxygenates DMSO to form a

new Lewis acid which then reacts with alcohols, equation 5. Two mechanisms below were postulated, equations 6 and 7.<sup>12</sup> In each case an active  $\text{NO}_2^+$  or  $\text{NO}_2\text{Cl}$  could be an electrophile for amide functionalization. Note that the structure in each has a structure similar to that of an anhydride like nitronium triflate. When the reaction was performed using  $\text{SOCl}_2$  as the acid chloride, a new product chemically different from 1-nitro-pyrrolidinone and starting material was observed. By mass spectrometry and NMR the product was tentatively assigned as the nitroso-amide **17**, equation 8.



## Conclusions

New, non-protic N-nitration methods were explored using nitronium triflate,  $\text{NO}_2\text{OSO}_2\text{CF}_3$  ( $\text{NO}_2\text{OTf}$ ), generated in situ from tetrabutylammonium nitrate,  $\text{Bu}_4\text{NNO}_3$ , and trifluoromethanesulfonic anhydride,  $(\text{CF}_3\text{SO}_2)_2\text{O}$  ( $\text{Tf}_2\text{O}$ ). 2-Pyrrolidinone, pyrrolidine, N-acetyl-pyrrolidine, N-acetyl-piperidine, N-acetyl-4-piperidinone, imidazolidinone, succinimide, and hydantoin were nitrated using these methods. Aqueous and non-aqueous workups gave yields ranging from 20-76%. Aqueous workup on 2-pyrrolidinone, pyrrolidine, N-acetyl-pyrrolidine, N-acetyl-piperidine, and N-acetyl-4-piperidinone containing one carbonyl involved in one amide linkage gave the best results. Non-aqueous workup on imidazolidinone, succinimide, and hydantoin gave the highest

yields. Investigations involving nitromethane as a solvent showed it did not increase the yield of products, and served only in a solubilizing capacity for reagents. Reactions using weak base, or excess starting material to allow it to participate as a base, did not improve the overall yields. The use of a quaternary ammonium salt nitrate source greatly improved the yields, and allowed all reactions to be run in pure dichloromethane. The reaction methods are significantly improved and eliminate the hazard potential, catalytic and solubilizing necessity and reaction workup requirements of nitromethane as a reaction solvent previously reported.

### Experimental Section

**General Comments:** All reactions were carried out under an atmosphere of dry nitrogen. Air and/or moisture sensitive reagents were handled by in a inert atmosphere dry box and flasks were kept under a positive nitrogen atmosphere. Methylene chloride was freshly distilled from the calcium hydride just prior to use. Tetrabutylammonium nitrate and trifluoromethanesulfonic anhydride was obtained from Aldrich and handled in a dry box. Acetic anhydride was obtained from Mallinckrodt. All other solvents and reagents were obtained from Aldrich, or reagent grade, and used without further purification. Reactions were monitored by thin layer chromatography on silica gel plates (E. Merck Kieselgel 60 F254) and column or filtration chromatography used silica gel (J. T. Baker, 60-200 mesh).

High-field NMR spectra were recorded on Bruker 300 spectrometer. Carbon-13 and proton NMR spectra were recorded at 75.469 MHz and 300.135 MHz respectively. Chemical shifts are reported in  $\delta$  units, part per million downfield, using TMS as the reference signal. Carbon-13 shifts are reported in  $\delta$  units using  $\text{CDCl}_3$  or DMSO as the solvent and/or reference signal. Infrared spectra were obtained using a Bio-Rad FTS-7 IR spectrometer with a SPC 3200 Data Station and run neat or as a Kubelka-Munk solid sample with KBr. EI mass spectra were recorded on a HP Model 5985 mass spectrometer, with an HP 1000 Data Station, operating at 70 eV. Melting points were taken using a Fisher Melt-Temp apparatus and are uncorrected.

**General Procedure for Nitrations Using  $\text{Bu}_4\text{NNO}_3$ /Triflic anhydride.:**  
Syntheses of 1-Nitropyrrolidinone (9), 1-Nitropyrrolidine (10), 1-Nitropiperidine (11), and 1-Nitro-4-piperidinone (12).

In a 100 mL three necked round bottom flask was placed 5.0 mmol (1.5224 g) of  $\text{Bu}_4\text{NNO}_3$  and to this added 30 mL anhydrous  $\text{CH}_2\text{Cl}_2$ . The system was placed under a positive nitrogen atmosphere and cooled to  $0^\circ\text{C}$  using a salt/ice bath. A 10 or 25 mL

pressure equalizing dropping funnel was charged with 5.0 mmol (1.4107 g) of triflic anhydride and attached to the reaction flask. The triflic anhydride was added dropwise to the reaction mixture and the resulting solution immediately turned yellow and slowly formed a pasty consistency with formation of a white solid. The mixture was stirred one hour at 0°C. In a dropping funnel was placed 5.0 mmol of substrate in approximately 3 mL of CH<sub>2</sub>Cl<sub>2</sub> and the solution added to the nitronium triflate reaction mixture. The reaction mixture was kept at the indicated temperature or warmed to room temperature (RT) and incubated for the indicated time. After the reaction was complete, 25 mL of 5% NaHCO<sub>3</sub> was added to quench the mixture and the solution stirred for 15-30 minutes. The reaction mixture was transferred to a separatory funnel and the CH<sub>2</sub>Cl<sub>2</sub> layer removed. The aqueous layer was extracted 1 x 25 mL CH<sub>2</sub>Cl<sub>2</sub> and the CH<sub>2</sub>Cl<sub>2</sub> layers combined. The organic layers were extracted 1 x 25 mL 5% NaHCO<sub>3</sub>, 1 x 25 mL saturated NaCl solution, dried with MgSO<sub>4</sub>, filtered and reduced in volume. A orange solid or oil was produced. The reaction mixture was chromatographed on a silica gel column, 1" x 5", using 400 mL of 1:3 ethyl acetate-hexane solvent and all fractions collected, combined and concentrated to give the product. Compounds **9**, **10**, **13**, **15**, and **16** gave proton and carbon spectra which agreed with reported literature values.<sup>1</sup> New and additional information for the compounds is given below.

**1-Nitropyrrolidinone (9):** IR (cm<sup>-1</sup>): 2980, 1774, 1766, 1559, 1554, 1276, 1229, 1143, 1021, 935, 839, 806, 760, 688, 629.

**1-Nitropyrrolidine (10):** IR (cm<sup>-1</sup>): 2961, 1731, 1629, 1494, 1371, 1308, 1261, 1026, 970, 910, 858, 763, 709.

**1-Nitropiperidine (11):** <sup>1</sup>H-NMR (δ, CDCl<sub>3</sub>): 3.86 (t, 4H), 1.72 (m, 4H), 1.60 (m, 2H); <sup>13</sup>C-NMR (δ, CDCl<sub>3</sub>): 49.10, 23.98, 22.57; IR (cm<sup>-1</sup>): 2940, 2857, 1642, 1515, 1442, 1387, 1326, 1276, 1240, 1063, 1012, 965, 854, 762; MW calcd. for C<sub>5</sub>H<sub>10</sub>N<sub>2</sub>O<sub>2</sub> 130.15 g/mol; Exact mass calcd. for C<sub>5</sub>H<sub>10</sub>N<sub>2</sub>O<sub>2</sub> 130.742206; Analysis calcd. C, 46.15; H, 7.75; N, 21.52; O, 24.59; Found.

**1-Nitro-4-piperidinone (12):** MP 74-76° C; <sup>1</sup>H-NMR (δ, CDCl<sub>3</sub>): 4.28 (t, 2H), 2.64 (t, 2H); <sup>13</sup>C-NMR (δ, CDCl<sub>3</sub>): 204.30, 46.42, 38.37; IR (cm<sup>-1</sup>): 3030, 2962, 2918, 1719, 1460, 1439, 1368, 1313, 1308, 1275, 1246, 1150, 1108, 1016, 974, 949, 828, 765, 744; MW calcd. for C<sub>5</sub>H<sub>8</sub>N<sub>2</sub>O<sub>3</sub> 144.13 g/mol; Exact mass calcd. for C<sub>5</sub>H<sub>8</sub>N<sub>2</sub>O<sub>3</sub> 144.0534855; Analysis calcd. C, 41.67; H, 5.60; N, 19.43; O, 33.30; Found.

**Synthesis of 1,3-Dinitro-2-oxotetrahydroimidazole (13).**

In a 100 mL three necked round bottom flask was placed 5.0 mmol of  $\text{Bu}_4\text{NNO}_3$  and to this added 30 mL anhydrous  $\text{CH}_2\text{Cl}_2$ . The system was placed under a positive nitrogen atmosphere and cooled to  $0^\circ\text{C}$  using a salt/ice bath. A 10 or 25 mL pressure equalizing dropping funnel was charged with 5.0 mmol of triflic anhydride and attached to the reaction flask. The triflic anhydride was added dropwise to the reaction mixture and the resulting solution immediately turned yellow and slowly formed a pasty consistency with formation of a white solid. The mixture was stirred one hour at  $0^\circ\text{C}$ . As a solid, 2.5 mmol of imidazolidinone was added to the nitronium triflate reaction mixture. The reaction mixture was warmed to room temperature (RT) and incubated for the indicated time. After the reaction was complete 25 mL of 5%  $\text{NaHCO}_3$  was added to quench the mixture and the solution stirred for 15-30 minutes. The reaction mixture was transferred to a separatory funnel and the  $\text{CH}_2\text{Cl}_2$  layer removed. The aqueous layer was extracted 1 x 25 mL  $\text{CH}_2\text{Cl}_2$  and the  $\text{CH}_2\text{Cl}_2$  layers combined. The organic layers were extracted 1 x 25 mL 5%  $\text{NaHCO}_3$ , 1 x 25 mL saturated  $\text{NaCl}$  solution, dried with  $\text{MgSO}_4$ , filtered and reduced in volume. An orange oil was produced. The reaction mixture was chromatographed on a silica gel column, 1" x 5", using 400 mL of 1:3 ethyl acetate-hexane solvent and all fractions collected, combined and concentrated to give the product.

(13): MP  $209-211^\circ\text{C}$ ;  $^1\text{H-NMR}$  ( $\delta$ , DMSO): 4.12; IR ( $\text{cm}^{-1}$ ): 2960, 2924, 2855, 2515, 2337, 2237, 1790, 1581, 1562, 1549, 1485, 1477, 1306, 1260, 1101, 1006, 948, 798, 754, 738, 715, 705.

Formation of **13** utilizing the general non-aqueous workup described below gave a yield of 32.6%.

#### **Synthesis of N-Nitrosuccinimide (15): General Non-aqueous workup.**

In a 100 mL three necked round bottom flask was placed 5.0 mmol of  $\text{Bu}_4\text{NNO}_3$  and to this added 30 mL anhydrous  $\text{CH}_2\text{Cl}_2$ . The system was placed under a positive nitrogen atmosphere and cooled to  $0^\circ\text{C}$  using a salt/ice bath. A 10 or 25 mL pressure equalizing dropping funnel was charged with 5.0 mmol of triflic anhydride and attached to the reaction flask. The triflic anhydride was added dropwise to the reaction mixture and the resulting solution immediately turned yellow and slowly formed a pasty consistency with formation of a white solid. The mixture was stirred one hour at  $0^\circ\text{C}$ . As a solid 5.0 mmol of finely powdered succinimide was added to the nitronium triflate reaction mixture. The reaction mixture was warmed to room temperature (RT) and incubated for the indicated time. After the reaction was complete the system was placed on a rotoevaporator and concentrated at room temperature under reduced pressure. The solid thus obtained was



dissolved with 1:1 ethyl acetate/hexane and layered onto a 1" x 5" silica gel column and eluted with 1:3 ethyl acetate/hexane. All fractions were collected and concentrated under reduced pressure to give a flaky cream colored solid. The melting point obtained was 82-84 °C, lower than the literature reported 89 °C. The proton and carbon NMR spectra agreed with the reported literature values.<sup>1</sup>

(15): MP 82-84° C; <sup>1</sup>H-NMR (δ, DMSO): 2.83; <sup>13</sup>C-NMR (δ, DMSO): 167.44, 26.86; IR (cm<sup>-1</sup>): 3552, 3005, 2952, 1742, 1590, 1423, 1407, 1302, 1266, 1163, 1048, 1003, 807, 732, 635.

### Synthesis of 1-Nitrohydantoin (16).

In a 100 mL three necked round bottom flask was placed 5.0 mmol of Bu<sub>4</sub>NNO<sub>3</sub> and to this added 30 mL anhydrous CH<sub>2</sub>Cl<sub>2</sub>. The system was placed under a positive nitrogen atmosphere and cooled to 0°C using a salt/ice bath. A 10 or 25 mL pressure equalizing dropping funnel was charged with 5.0 mmol of triflic anhydride and attached to the reaction flask. The triflic anhydride was added dropwise to the reaction mixture and the resulting solution immediately turned yellow and slowly formed a pasty consistency with formation of a white solid. The mixture was stirred one hour at 0°C. As a solid 2.5 mmol of finely powdered hydantoin was added to the nitronium triflate reaction mixture. The reaction mixture was warmed to room temperature (RT) and incubated for the indicated time. After the reaction was complete the system was placed on a rotoevaporator and concentrated at room temperature under reduced pressure. The solid thus obtained was dissolved with 1:1 ethyl acetate/hexane and layered onto a 1" x 5" silica gel column and eluted with 1:3 ethyl acetate/hexane. All fractions were collected and concentrated under reduced pressure to give a pale yellow solid. On allowing the solid to stand in a closed round bottom flask overnight, a brown orange gas developed which could be easily removed by flushing with nitrogen. The remaining white solid gave pure 16.

(16): MP 171-173° C; <sup>1</sup>H-NMR (δ, DMSO): 11.99 (s, 1H), 4.63 (s, 2H); <sup>13</sup>C-NMR (δ, DMSO): 166.02, 149.52, 51.92.

### Synthesis of N-Acetyl-pyrrolidine (methyl pyrrolidinyll ketone) (3).

The synthesis of N-acetyl-pyrrolidine was a modified procedure of Woodward.<sup>13</sup> Additional procedures consulted were those used to synthesize HMX<sup>1</sup> and RDX,<sup>2</sup> however the N-acetyl-pyrrolidine was expected to be a liquid, limiting the isolation techniques utilized for HMX and RDX. To a 500 mL round bottom flask containing 20.40 g (0.198 mol) of acetic anhydride in 100 mL CH<sub>2</sub>Cl<sub>2</sub>, cooled to 0° C, was added dropwise a

solution of 14.22 g (0.199 mol) of pyrrolidine in 50 mL  $\text{CH}_2\text{Cl}_2$  over 15 minutes. An exothermic reaction occurred. The system was stirred 16 hours at room temperature. The system was placed in a large beaker, with magnetic stirring, and to it slowly added 100 mL 5%  $\text{NaHCO}_3$ . The solution evolved  $\text{CO}_2$  and was further quenched with solid  $\text{NaHCO}_3$  until no evolution of  $\text{CO}_2$  occurred. The system was placed in a separatory funnel and 200 mL  $\text{H}_2\text{O}$  added, the system shaken, and the  $\text{CH}_2\text{Cl}_2$  removed. The aqueous layer was extracted 2 x 150 mL 5%  $\text{NaHCO}_3$ , 2 x 500 mL 1N  $\text{HCl}$ , 1 x 500 mL saturated  $\text{NaCl}$  solution, dried with  $\text{MgSO}_4$ , filtered and concentrated under reduced pressure to give a colorless oil, 1.299 g (11.5 mmol), 5.8% yield. Note: This procedure gave a sufficient quantity of product for initial runs however it does not represent a feasible route for synthesis.

(3):  $^1\text{H}$ -NMR ( $\delta$ ,  $\text{CDCl}_3$ ): 3.44 (m, 4H), 2.04 (s, 2H), 1.84-2.09 (m, 4H);  $^{13}\text{C}$ -NMR ( $\delta$ ,  $\text{CDCl}_3$ ): 168.62, 46.91, 45.01, 25.63, 24.11; IR ( $\text{cm}^{-1}$ ): 3452, 2970, 2872, 1626, 1452, 1427, 1343, 1225, 1194, 1020, 986.

#### Non-aqueous workup: Nitration of 2-pyrrolidinone (9).

An identical reaction was repeated as for 2-pyrrolidinone until the initiation of the workup. Following incubation for the 1 hour, the system was placed on a rotoevaporator and concentrated at room temperature under reduced pressure to give an orange oil. The oil thus obtained was dissolved with 1:1 ethyl acetate/hexane and layered onto a 1" x 5" silica gel column and eluted, with 1:3 ethyl acetate/hexane, a pale yellow band which on concentration under reduced pressure gave a yellow oil. This oil slowly solidified. The solid was insoluble in  $\text{CDCl}_3$  but soluble in  $\text{DMSO}-d_6$ . The spectra was too complicated for assignment and the major signals observed are listed below.

$^1\text{H}$ -NMR ( $\delta$ ,  $\text{DMSO}$ ): 11.96 (br), 6.08 (v br), 5.11 (m), 4.27 (t), 4.04 (m), 3.44 (m), 2.43 (m), 2.29 (m), 2.16 (m), 1.76 (m), 1.22 (m), 1.11 (m);  $^{13}\text{C}$ -NMR ( $\delta$ ,  $\text{DMSO}$ ): 174.04, 117.88, 73.21, 67.24, 63.60, 60.14, 44.55, 30.86, 27.56, 21.96.

#### Acknowledgements

We gratefully acknowledge support by the Air Force Office of Scientific Research and the Air Force Summer Faculty Research Program, administered by the Research and Development Laboratory for this work. C. M. A. wishes to personally thank Major Scott A. Shackelford, and the military and civilian personnel at the F. S. Seiler Research Laboratory for their stimulating and enlightening discussions and direction on the many aspects of research performed at the facility.

## References:

- 1) Shackelford, S. A.; Goshgarian, B. B.; Chapman, R. D.; Askins, R. E.; Flanigan, D. A.; Rodgers, P. N. *Propellants, Explosives, Pyrotechnics* **1989**, *14*, 93-102, and references cited therein.
- 2) Rodgers, S. L.; Coolidge, M. B.; Lauderdale, W. J.; Shackelford, S. *Thermochemical Acta* **1991**, *177*, 151-168.
- 3) Shackelford, S. A., "Mechanistic Investigations of Condensed Phase Energetic Material Decomposition Processes Using the Kinetic Deuterium Isotope Effect", Bulusu, S. N., Ed. *Chemistry and Physics of Energetic Materials*, Kluwer Academic Publisher, Netherlands, 1990, pp. 413-432.
- 4) Shackelford, S. A., "Mechanistic Relationships of the Decomposition Process to Combustion and Explosive Events from Kinetic Deuterium Isotope Effect Investigations", Bulusu, S. N., Ed. *Chemistry and Physics of Energetic Materials*, Kluwer Academic Publisher, Netherlands, 1990, pp. 433-456.
- 5) Suri, S. C.; Chapman, R. D. *Synthesis* **1988**, 743-745.
- 6) Feuer, Henry and Nielsen, Arnold T., Eds. "Nitro Compounds: Recent Advances in Synthesis and Chemistry: Organic Nitro Chemistry Series", VCH Publishers, Inc., NY, NY., 1990.
- 7) Barco, A.; Benetti, S.; Pollini, G. P.; Spalluto, G.; Zanirato, V. *Tetrahedron Lett.* **1991**, *32*(22), 2517-2520.
- 8) Shackelford, S. A.; Sharts, C., unpublished results.
- 9) Butyrolactone:  $^1\text{H-NMR}$  (Satler No. 14002):  $\delta$  1.9-2.60, 4.29;  $^{13}\text{C-NMR}$  (Bassler and Silverstein, 4th Ed.)  $\delta$  177.9, 68.6, 27.7, 22.2.
- 10) Alcoholysis of some N-nitro derivatives have been reported: Koslova, I. K.; Luk'yanov, O. A.; Tartakovskii, V. A. *Izv. Akad. Nauk. SSSR Ser. Khim.* **1981**, *11*, 2556-2563. b) Koslova, I. K.; Luk'yanov, O. A.; Tartakovskii, V. A. *Izv. Akad. Nauk. SSSR Ser. Khim.* **1981**, *11*, 2563-2571.
- 11) Chiba, K.; Tagaya, H.; Karasu, M.; Ono, T.; Hashimoto, K.; Moriwaki, Y. *Bull. Chem. Soc. Jpn.* **1991**, *64*, 966-970.
- 12) March, J. "Advanced Organic Chemistry", 3rd Ed. J. Wiley and Sons, Inc. NY, NY, 1985, pp. 468-470.
- 13) Woodward, R. B.; von E. Doering, W. J. *Am. Chem. Soc.* **1945**, *67*, 860-874.

**THERMAL DECOMPOSITION OF TNT, NTO, AND THEIR MIXTURES  
VIA ISOTHERMAL DIFFERENTIAL SCANNING CALORIMETRY**

**DR. GARY S. BUCKLEY**

**ABSTRACT**

The thermal decomposition of 2,4,6-trinitrotoluene (TNT), 3-nitro-1,2,4-triazol-5-one (NTO), their deuterated analogs, and mixtures thereof have been studied using isothermal differential scanning calorimetry. Primary isotope effects previously noted for both TNT and NTO decompositions were diluted or altered, probably due to the different sample pan configuration used here. TNT decomposition appears to be catalyzed by the presence of Viton rubber, while the NTO decomposition appears to be catalyzed by confinement of its product gasses. Mixture studies indicate that a 2:1 mole ratio of TNT to NTO is sufficient to cause the accelerated decomposition of TNT and consume all of the NTO present in the process.

**Introduction**

Although NTO was first synthesized in 1905<sup>1</sup>, interest in its use as an explosive has just recently been shown<sup>2</sup>. NTO does not appear to suffer the sensitivity problems associated with its more famous counterparts such as TNT, RDX, and HMX. A less sensitive explosive currently in use, TATB, does not have the energetic performance of RDX or HMX. Thus a need exists for the development of an explosive that is more safely handled than TNT, HMX, and RDX but that will also deliver the explosive characteristics. NTO could potentially fill such a role.

Mixtures of TNT and NTO are currently under investigation as potential explosives. Though the decomposition of TNT has been extensively studied<sup>3</sup>, very little work has been directed toward studying the decomposition of NTO and its mixtures with TNT<sup>4</sup>. Of particular interest in the study of mixtures is an understanding of the mechanism of the mutual decompositions and the effects of varying compositions of TNT and NTO.

The primary tool to be used to understand the mechanism of the decomposition is the kinetic deuterium isotope effect (KDIE)<sup>5</sup>. Briefly, substitution of a deuterium for a hydrogen in a compound may affect the rate constants observed from the reactions of that compound. Isotope effects are broken into three categories - primary, secondary, and inverse. A primary isotope effect is observed if the substituted bond is broken in the rate limiting step of the reaction. A primary isotope effect leads to a rate constant ratio of hydrogenated to deuterated ( $k_H/k_D$ ) compound that exceeds 1.35. A secondary isotope effect is observed if a bond other than the substituted bond is broken in the rate limiting step of the reaction. A secondary isotope effect leads to a  $k_H/k_D$  ratio in the range of 1.00-1.34, depending upon the distance of the bond broken in the transition state from the substitution. The third isotope effect is referred to as an inverse isotope effect. In this case, the  $k_H/k_D$  ratio is less than one.

The KDIE has been used successfully in the past to study the thermal decomposition of TNT, HMX, RDX, and TATB<sup>6</sup>. In this work, the thermal

decompositions of NTO and NTO-D2 are studied in an effort to better understand their mechanisms. Additionally, mixtures of TNT and NTO and their deuterated analogs are studied to evaluate any potential catalytic effect of each compound on the other as well as to better understand the mechanism of their thermal decomposition.

#### Experimental Section and Data Analysis

Isothermal differential scanning calorimeter experiments were conducted on a Perkin-Elmer DSC-7 system connected to a Perkin Elmer 7500 Data Station. Perkin-Elmer large volume capsules (PE part no. 0319-0021) were used for all experiments. Though some experiments were conducted without the O-ring seal, the numerical data reported here comes from capsules that were sealed with O-rings. Samples were loaded into the DSC head at 50 °C and the temperature was increased at the rate of 200 °C/min to the selected isothermal temperature. All sample sizes were in the range of 1.75-2.15 mg.

Samples of TNT, TNT-D3, NTO and NTO-D2 were prepared during the summer of 1990 by Ted Burkey<sup>7</sup> and had been stored in a dessicator. A melting point determination of the NTO and TNT-D3 indicated that the material was substantially unchanged over one year.

Mixtures that were studied were made by dissolving the desired quantities of each component in acetone. Once dissolution was complete, a stream of dry nitrogen was passed into the flask until the acetone was no longer visually evident. The flask was then placed in a vacuum dessicator for approximately one hour to remove the last traces

of acetone.

Data from the 7500 Data Station were fed into an IBM compatible PC machine in a single byte hexadecimal format. A macro program was written in QUATRO PRO to convert the data to decimal format and allow for further calculation. Due to memory constraints in the PC, some files needed to be reduced in size. This was accomplished by reading in every other point or every third point depending upon the number of data points in the DSC file.

## Results

### Sample Pan Configuration

A wide variety of sample pan configurations are available for use with the Perkin-Elmer DSC-7. Variables include material of construction, volume, and pressure containment or release. Past work<sup>7</sup> had indicated that stainless steel large volume capsules would work in this study, and that it would be beneficial to seal such pans with a Viton O-ring to contain gaseous products.

There are at least three drawbacks to sealing the vessel. The first is that the various experiments will no longer be conducted at the same measurable pressure. Measurements of mass loss after experiments indicate that decomposition processes for NTO and TNT are considerably different, with the NTO converting approximately  $59 \pm 8\%$  of its mass to gaseous (SATP) products. TNT, on the other hand, only converts about  $14 \pm 1\%$  to gaseous products. Thus the pressure contained in a sealed pan at the end of the decomposition could be considerably different for

the two materials.

A second drawback to sealing the vessels is that the product gases contained during decomposition could have a catalytic effect on the process. This catalytic effect could alter the mechanism of decomposition or mask the true decomposition kinetics in the system.

A third drawback to sealing the vessels is that the O-ring itself could catalyze the decomposition process. NTO decomposes in the solid state at approximately 262 °C (visual verification in melting point block), so it is unlikely that the O-ring, which is not in contact with the sample, would accelerate the NTO decomposition process. TNT, on the other hand, melts at 80 °C and apparently condenses to a large extent on the ceiling of the pan (visual verification after experiments) and could thus have an opportunity to come in contact with the O-ring.

In order to investigate these effects, experiments were conducted using cells that were sealed, not sealed, or only partially sealed. For the partially sealed cells, approximately one-fourth of the O-ring was removed and the stainless steel pan was "sealed" with the remaining three-fourths O-ring. The peak time for the TNT at 257 °C without O-ring was considerably delayed compared to the sealed vessel (30 min. compared to 15 min.). In addition, the unsealed vessel gave a less well defined and reproducible exotherm than the sealed. Mass loss measurements indicate that the stainless steel pans do not seal consistently if the O-ring is omitted. Inclusion of a three-fourths



O-ring in the vessel gave essentially two exotherms - a sharp one immediately upon hitting 257 °C and a much smaller delayed one that corresponds approximately to the large exotherm of the sealed system. Addition of O-ring bits to the sample pan without an O-ring seal led to results that were very similar to the first exotherm in the partially sealed system. It appears that the TNT decomposition is catalyzed by contact with the O-ring.

Sealing of the vessel containing NTO with a Viton O-ring also apparently catalyzes its thermal decomposition relative to the unsealed vessel. Containment of the pressure leads to a sharper more reproducible exotherm. Mass loss measurements indicate that virtually all of the gaseous products escape during the run if the O-ring seal is omitted. A partially sealed vessel results in a thermogram quite similar to the unsealed vessel. Addition of O-ring bits to the sample pan produced no visible exotherm after 40 min at 235 °C in an unsealed vessel, while NTO decomposed in 20 minutes in an O-ring sealed vessel at the same temperature. This indicates that the containment of the product gasses or the increased pressure is more likely catalyzing the NTO decomposition than the presence of the O-ring.

The experiments conducted here were conducted in vessels sealed with O-rings due to the better reproducibility and the large amount of gaseous products formed, particularly in the case of NTO. Activation parameters determined will reflect the catalyzed process and may not be readily transferable to the uncatalyzed processes. However, the

relative activation parameters may shed considerable light on the interaction of NTO and TNT and their deuterated analogs during the thermal decomposition process.

#### General Description of Analysis

Figure 1 illustrates the general appearance of all DSC thermograms obtained in this study. There are two exotherms evident - a small one early in time that generally initiates immediately upon reaching the chosen temperature and a second larger one that follows.

Though there could be some interesting aspects to the first exotherm, temperatures used in this study were not low enough to observe the whole exotherm. One way of approximating kinetic data from the first peak is to record an induction time. This corresponds to the high heat flow point between the first and second exotherm (point A in Figure 1). This induction time corresponds roughly to the time of completion of the first exothermic process. This set of data is referred to as induction time data in the following discussion.

The second exotherm allows for a more complete kinetic analysis. The QUATTRO PRO program produced the three plots given in Figure 2 from which the quantitative analysis proceeded.

The first plot includes the data, the selected baseline and the natural logarithm of the baseline minus the heat flow.

The second plot compares an autocatalytic treatment of the selected peak with a first order treatment. Rogers<sup>8</sup> has shown that, for a first order reaction, the  $\ln b$ , where  $b$  is the heat flow, is linearly related to the time. Implicit in this derivation is also the fact that for a first order reaction  $-\ln(1-\alpha)$  (where  $\alpha$  is the degree of conversion) must be linearly related to time. The first two plots allow one to compare the linearity of  $\ln b$  (from the first plot) with the linearity of  $-\ln(1-\alpha)$  (from the second plot). If both sets of data are not linear in the same region, the reaction is most likely not first order. In this study, there did not appear to be substantial first order behavior in any of the experiments.

The second curve of the second plot illustrates the behavior of  $1-\alpha$  versus time. In autocatalytic reaction, this plot will assume a characteristic decreasing sigmoidal shape.

The third plot of Figure 2 allows one to calculate the autocatalytic rate constant. An autocatalytic reaction will render a significant linear portion in this plot<sup>9</sup> with a slope of  $k$ . All of the experiments conducted here were treated with autocatalytic kinetics. In every case, the linear portion extended at least from  $(1-\alpha)$  values of 0.06 to 0.17 which corresponds to  $\alpha$  of 0.065 to 0.22.

#### TNT and TNT-D3:

Figure 3 gives the Arrhenius plots for TNT and TNT-D3 where  $k$  comes from the autocatalytic treatment (Figure 3a) and where  $k$  is simply the reciprocal of the induction time (Figure 3b). Autocatalytic treatment

of the data leads to an activation energy ( $E_A$ ) for TNT decomposition of  $27.5 \pm 0.8$  kcal/mol and for TNT-D3 of  $23 \pm 2$  kcal/mol. Omission of the two higher temperature points from TNT-D3 leads to an  $E_A$  of  $27 \pm 1$  kcal/mol, in close agreement with the TNT. Comparison of the induction time activation energies leads to similar activation energies again,  $22.0 \pm 0.6$  kcal/mol for TNT and  $22 \pm 1$  kcal/mol for TNT-D3. Activation energies obtained from the autocatalytic treatment are very similar to those found in past work conducted with unsealed aluminum pans.<sup>3</sup>

The average isotopic ratio obtained in this work is  $1.11 \pm 0.05$  for autocatalytic treatment and  $1.11 \pm 0.09$  for induction time treatment over the temperature range studied (245-265°C). This contrasts quite sharply with values found earlier<sup>3</sup> of  $1.35 \pm 0.02$  and  $1.6 \pm 0.02$ , respectively. The major differences in the studies center on the sample pan configuration. The aluminum pans used in the earlier study will not contain pressure above ambient, while the stainless steel pans used in this study will contain up to 24 atmospheres when sealed with a rubber O-ring. The O-ring itself introduces another factor as noted earlier, the possibility of catalytic action.

As was noted earlier, the TNT decomposition seems to be catalyzed by the presence of the O-ring. A question arises as to whether the catalytic effect of the O-rings dilutes or even removes the isotope effect. A comparison of the past work and current work suggests that there is such an effect. A comparison of TNT and TNT-D3 runs without

O-rings at 250 °C gives an induction time ratio of 3.12, much different than the values obtained when O-rings are present. Mass loss measurements indicate that loss of product gasses is not a problem even in the absence of the O-ring. Presence of the O-ring apparently alters the mechanism or speeds the rate enough that the isotope effect either becomes secondary in nature or is diluted in magnitude.

#### NTO and NTO-D2

Figure 4 gives the Arrhenius plots for NTO and NTO-D2 where  $k$  comes from the autocatalytic treatment (Figure 4a) and where  $k$  is simply the reciprocal of the induction time (Figure 4b). The autocatalytic plots are scattered, particularly for NTO-D2. It is difficult, based upon the data presented, to arrive at a reasonable linear fit. The induction time plot is, however, much improved over the autocatalytic treatment and leads to activation energies of  $47.5 \pm 1.1$  and  $53.6 \pm 3.5$  kcal/mol for NTO and NTO-D2, respectively. These values are considerably lower than earlier work<sup>4</sup> which produced activation energies of  $87.5 \pm 1.8$  and  $88.4 \pm 1.2$  kcal/mol for NTO and NTO-D2, respectively. There is the possibility that the stainless steel catalyzes the NTO decomposition relative to the glass used in the other study or that confinement of product gasses is affecting the decomposition.

Isotopic ratios obtained in this work seem to decrease rather dramatically as temperature increases over the limited temperature range investigated (230-243 °C). Due to the substantially higher activation energy of NTO over TNT, the temperature region over which useful data may be obtained is limited. At low temperature (220-230

$^{\circ}\text{C}$ ), thermograms are erratic and not generally reproducible. At higher temperatures (above  $245^{\circ}\text{C}$ ), the reaction initiates before the DSC has stabilized. It is difficult to say from this work unequivocally whether the isotope effect is primary or secondary for NTO decomposition under these conditions.

#### TNT/NTO Mixtures

Figure 5 gives Arrhenius plots for the thermal decomposition of TNT and NTO mixtures where data is again extracted from the second major exotherm. There is apparently no large effect due to the isotopic substitutions indicated in the figure. Activation energies for the 36 mol % TNT samples range from  $60 \pm 2$  kcal/mol for TNT/NTO to  $65 \pm 1$  kcal/mol for the TNT-D<sub>3</sub>/NTO-D<sub>2</sub> mixtures. Isotopic ratios range from 1.42 at  $220^{\circ}\text{C}$  to 1.18 at  $235^{\circ}\text{C}$ . The activation energy for the decomposition is in the neighborhood of the NTO activation energy, and the peak area (average 1120 J/g) is in the vicinity expected for a 36 mol% TNT mixture (885 J/g). The second exotherm in these mixtures may correspond to the decomposition of excess NTO. The TNT decomposition apparently corresponds to the first exotherm observed.

It is of interest to determine at which composition of mixture the second exotherm is no longer evident. Data taken with a 1.9 mol % NTO mixture suggest that NTO decomposition at this level is either not within instrumental limits of detection or that the small amount of NTO is decomposed in the early exotherm. NTO levels were increased to 17, 33, 50, and 64 mol %. Figure 6 shows the results of the ensuing thermograms at  $235^{\circ}\text{C}$ . It appears that the NTO concentration beyond

the 33 mol % level results in excess NTO remaining after catalyzing the TNT decomposition.

### Conclusions

Isotope effects previously seen in TNT decompositions using IDSC were not as prominent in this work. Apparently the presence of the Viton O-ring catalyzes the TNT decomposition and dilutes the isotope effect or alters the mechanism sufficiently to change the rate determining step.

The thermal decomposition of NTO can be followed by IDSC over only a very narrow range of temperature due to its higher activation energy than that of TNT. Results suggest a possible isotope effect that is strongly temperature dependent in O-ring sealed stainless steel pans. The decomposition of the NTO is catalyzed by the confinement of its products gasses.

Mixture studies of TNT and NTO suggest that a 33 mol % mixture of NTO is sufficient to catalyze the decomposition of TNT. Levels of NTO in excess of that result in excess NTO remaining after TNT decomposition. Computer calculations have indicated that the energies of the two N-H bonds on NTO are similar<sup>10</sup>, so it may be possible that two TNT molecules can attack each NTO.

A method that could be used to better study the thermal decomposition of NTO is thermogravimetric analysis (TGA). The NTO decomposition produces about 60% of its mass as gaseous products at SATP, but the

decomposition apparently occurs in the solid phase. Also, since the TGA does not have the settling time that is a problem in DSC, the temperature range to be studied could perhaps be expanded. TGA would also allow product gases to be swept away, removing the possibility of catalytic action of these species.

#### REFERENCES

1. Manchot, W.; Noll, R. *Justus Liebigs Ann. Chem.* 1905, 1, 343.
2. Lee, K. Y.; Chapman, L. B.; Coburn, M. D. *J. Energetic Matls.* 1987, 5, 27.
3. Shackelford, S. A.; Beckmann, J. W.; Wilkes, J. S. *J. Org. Chem.* 1977, 42, 4201.
4. Menapace, J. A.; Marlin, J. E.; Bruss, D. A.; Dascher, R. V. *J. Phys. Chem.* 1991.
5. For example:  
  
Allinger, N. L.; Cava, M. P.; DeJongh, D. C.; Johnson, C. R.; Lebel, N. A.; Stevens, C. L. *Organic Chemistry*; Worth Publishers, Inc.: 1971; pp.300-301.  
  
Bigeleisen, J.; Wolfsberg, M. in *Advances in Chemical Physics* 1958, 1, pp.15-31.  
  
Lowery, T. H.; Richardson, K. S. *Mechanism and Theory in Organic Chemistry*; Harper and Row: New York, 1981; pp. 206-208, 211-212.
6. TNT: See reference 3.  
  
HMX: Shackelford, S. A.; Coolidge, M. B.; Goshgarian, B. B.; Loving, B. A.; Rogers, R. N.; Janney, J. L.; Ebinger, M.H. *J. Phys. Chem.* 1985, 89,3118.  
  
RDX: Shackelford, S. A.; Rodgers, S. L.; Coolidge, M. B. *CPIA Publ.* 412, II, 1984, 615.  
  
TATB: Rogers, R. N.; Janney, J. L.; Ebinger, M. H. *Thermochim. Acta* 1982, 59, 287.
7. Burkey, T. J. Final Report of 1990 USAF-UES Summer Faculty Research Program.



8. Rogers, R. N. *Anal. Chem.* 1972, 44, 1336.
9. Rogers, R. N. *RCM Report A-04-87* 1987, 15.
10. Ritchie, J. P. J. *Org. Chem.* 1989, 54, 3553.

DSC Data File: gbl10  
 Sample Weight: 1.860 mg  
 Mon Jun 03 10:52:23 1991  
 TNT - SS Pens with O-rings

PERKIN-ELMER

7 Series Thermal Analysis System

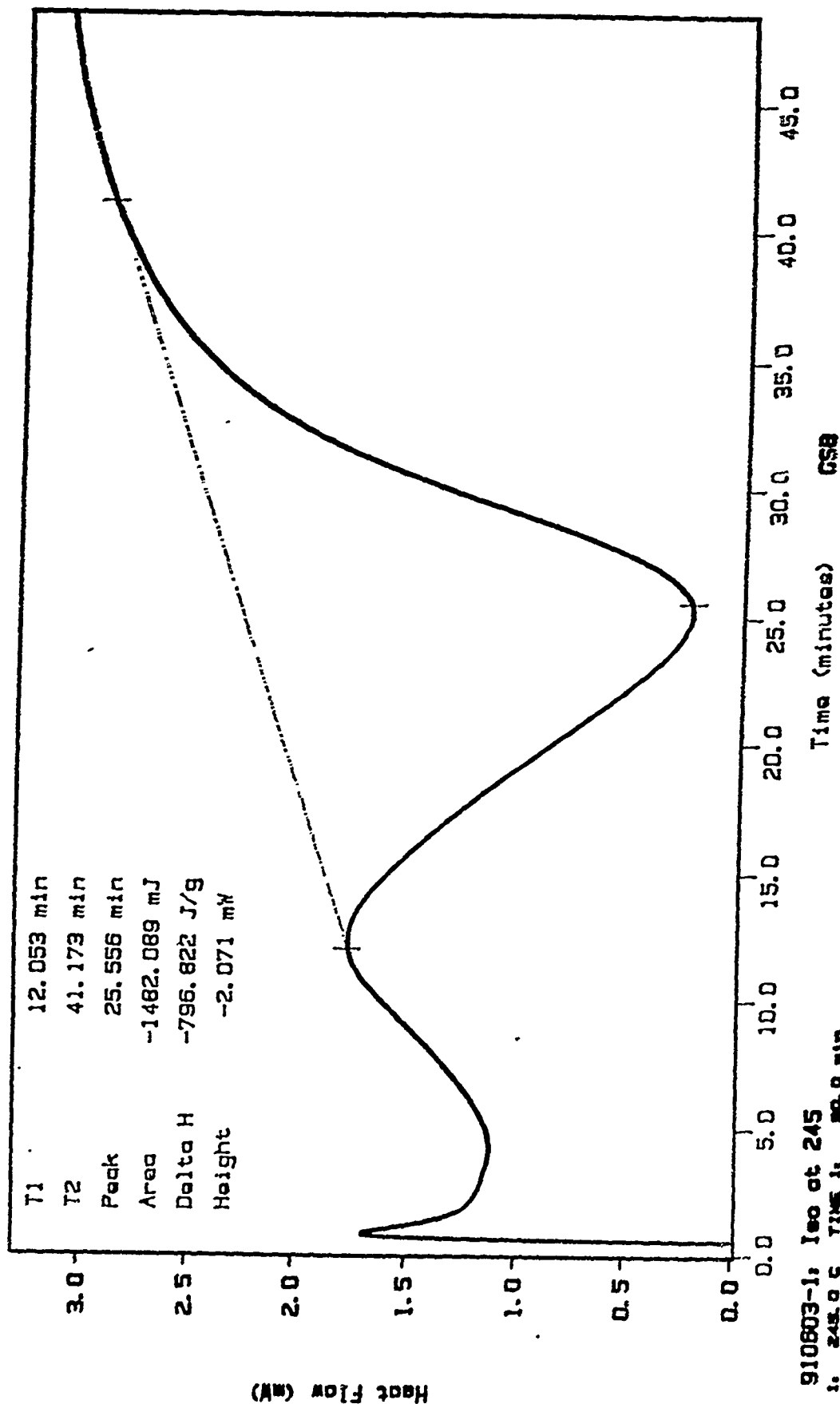


Figure 1

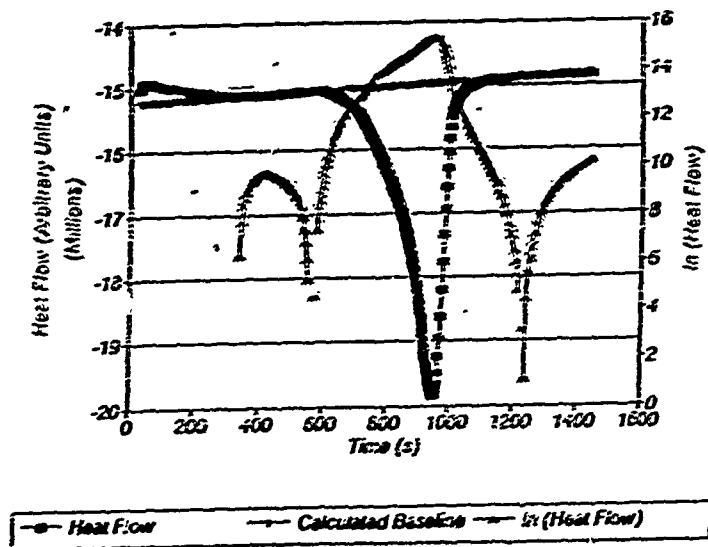


Figure 2a

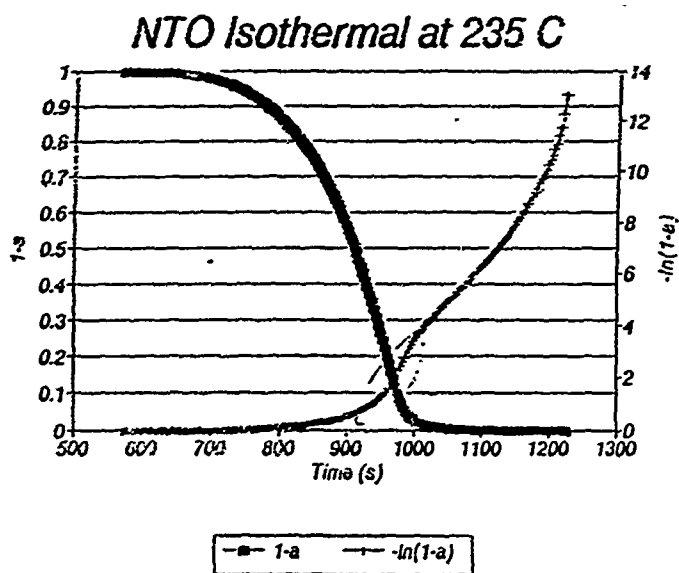
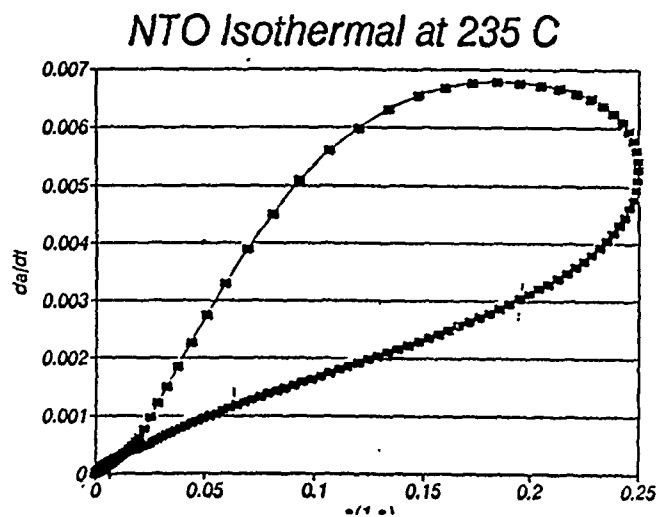


Figure 2b



## Autocatalytic Plot for TNT and TNT-D3

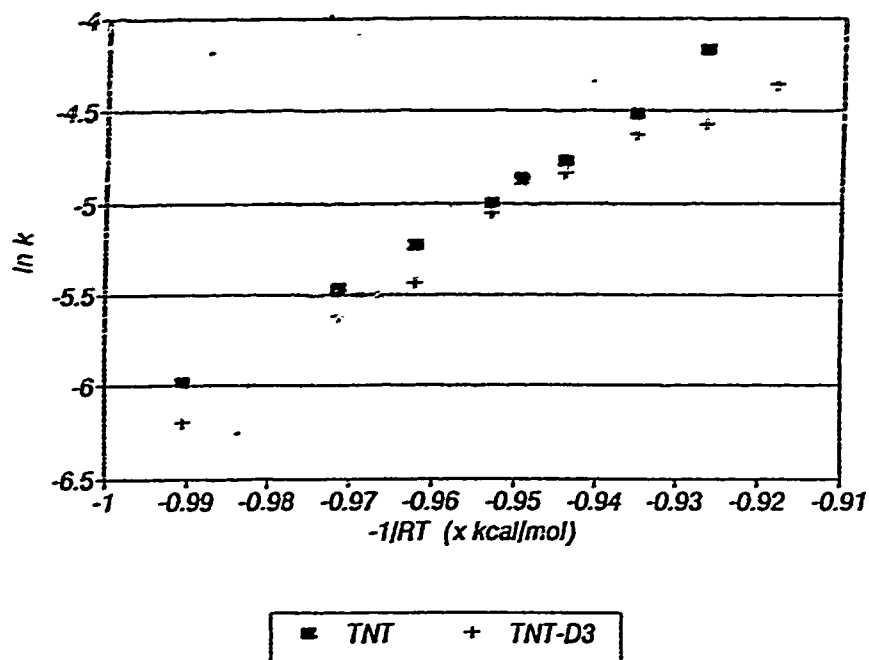
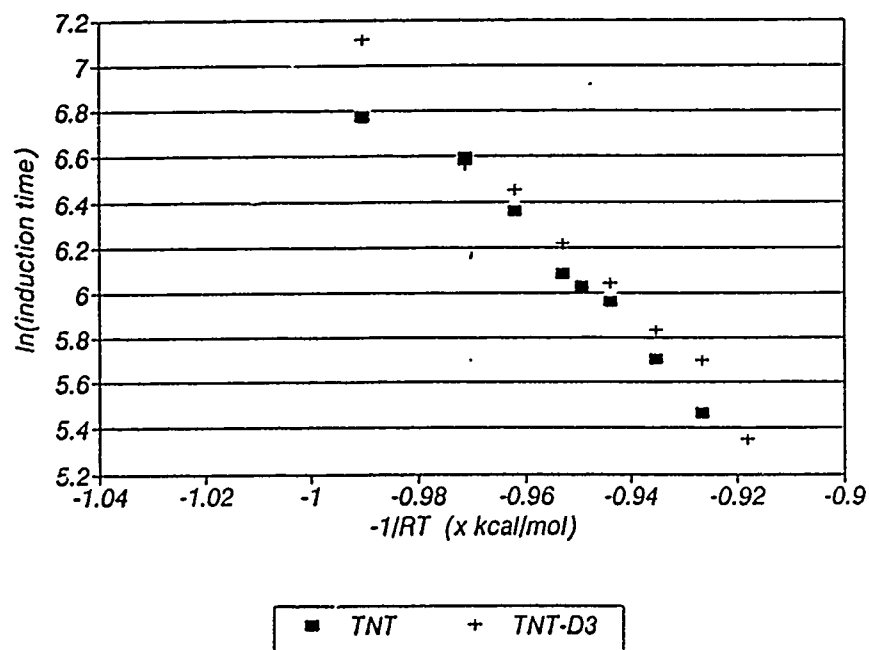


Figure 3a

## Induction Time Plot for TNT and TNT-D3



## Autocatalytic Plot for NTO and NTO-D2

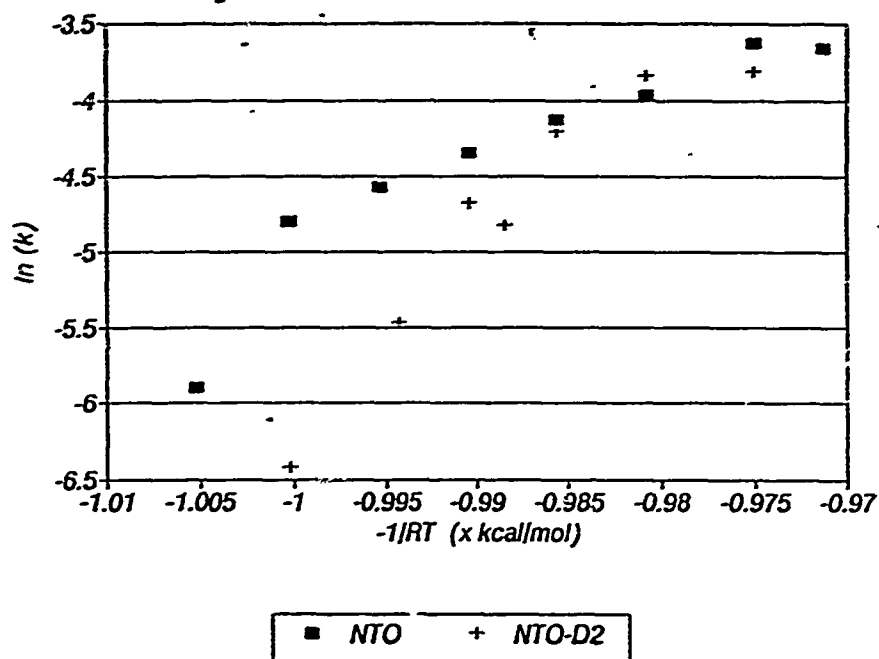
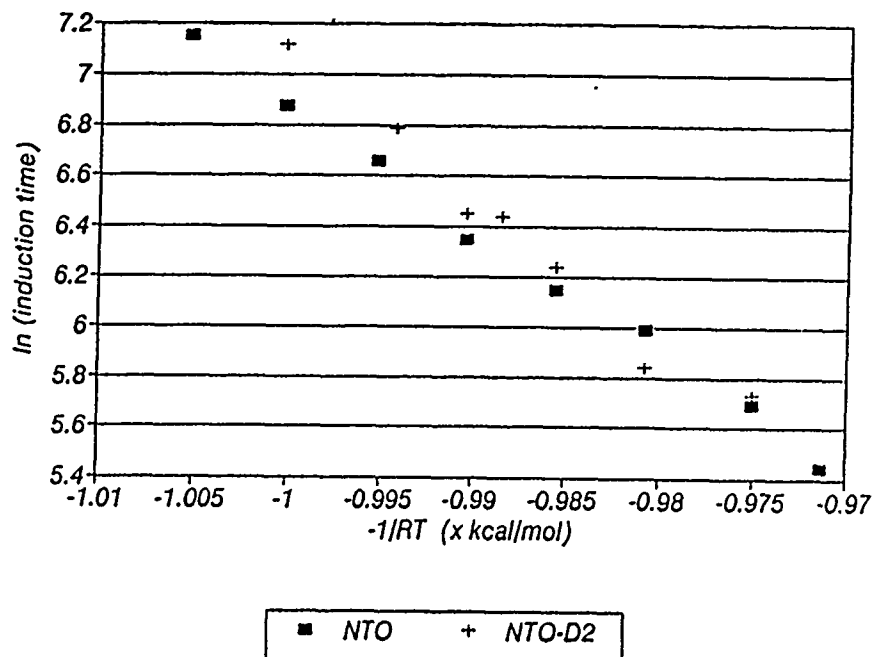


Figure 4a

## Induction Time Plot for NTO and NTO-D2



# Autocatalytic Plot for TNT/NTO Mixtures

## 50 wt. % Mixtures

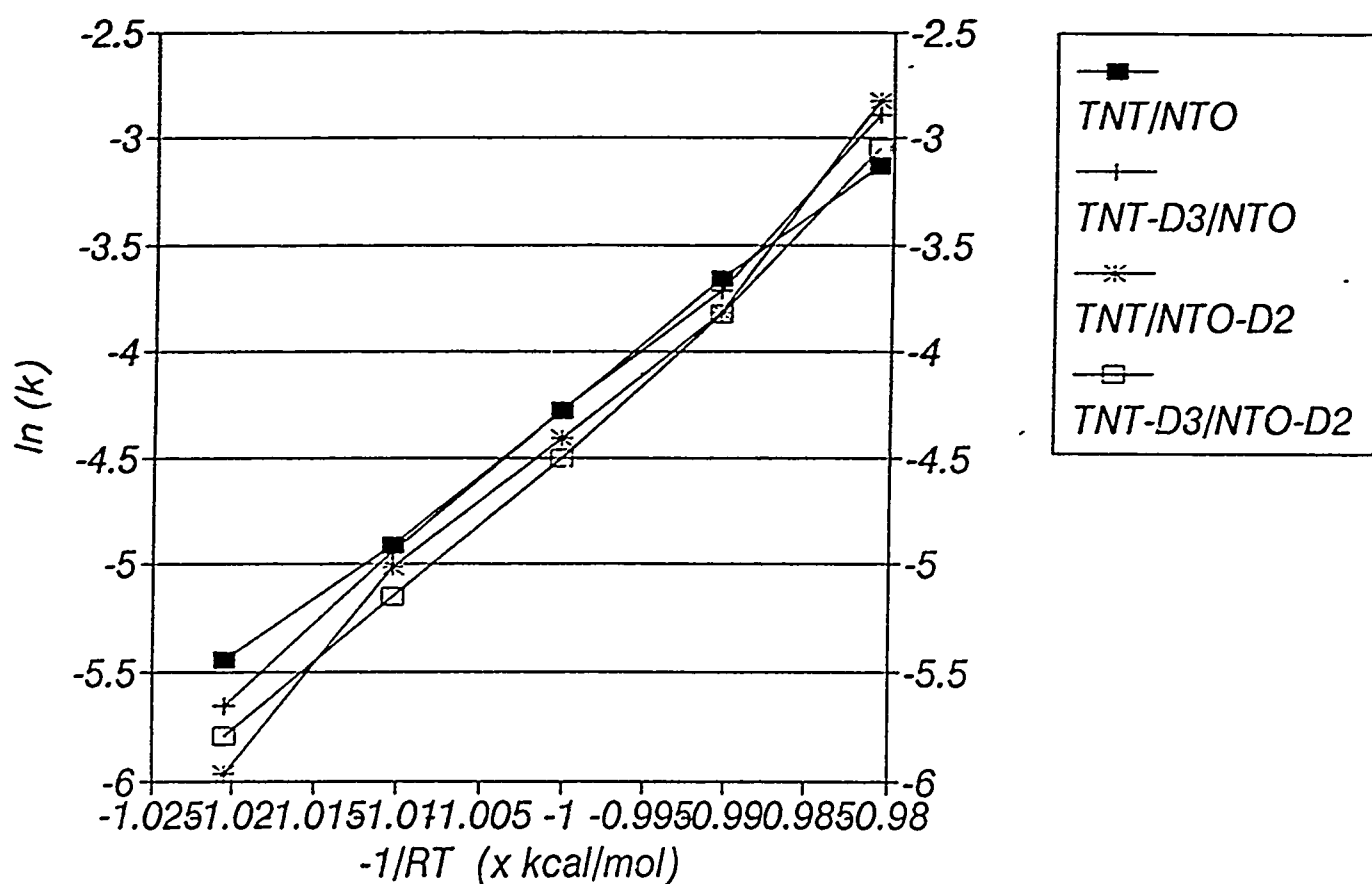
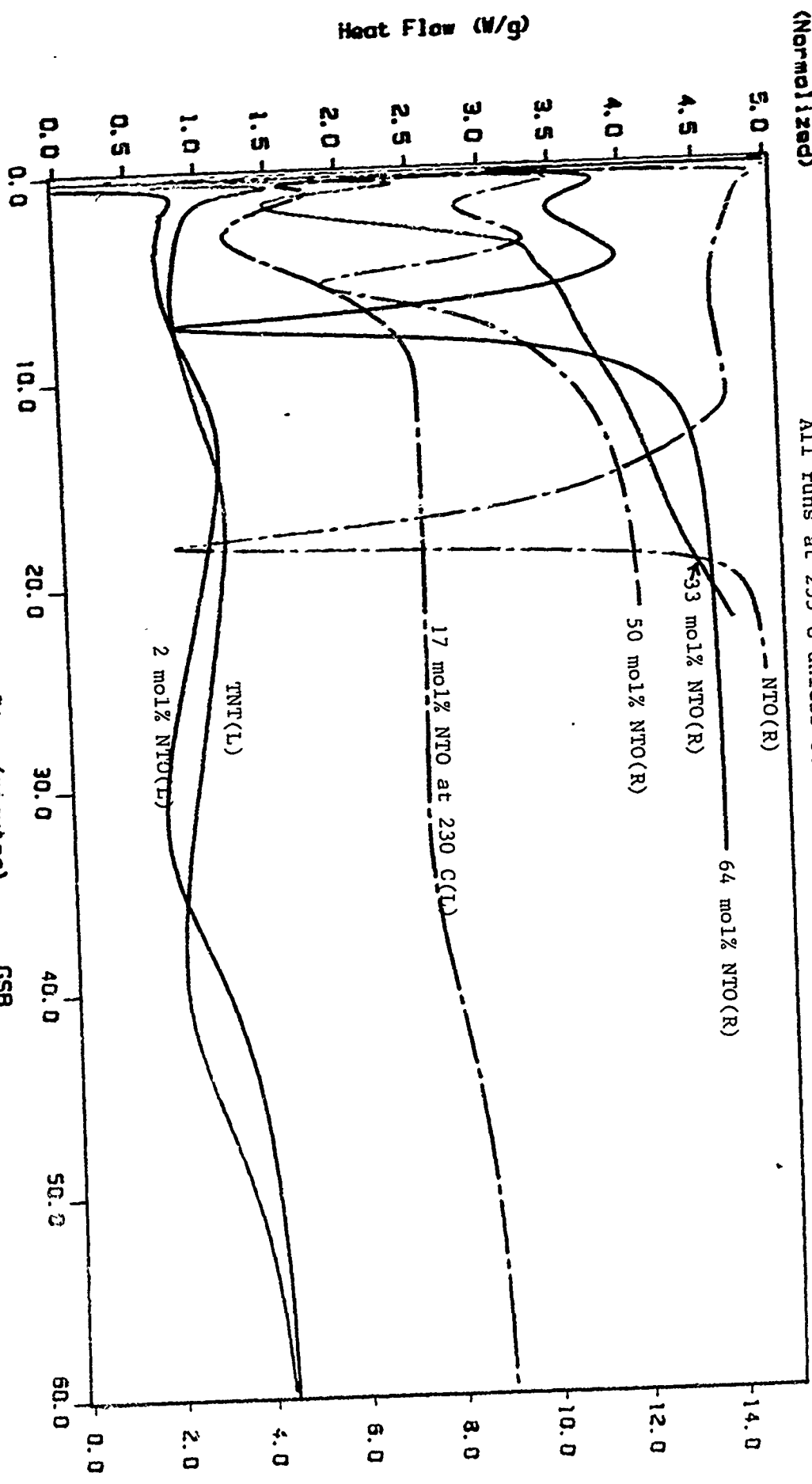


Figure 5

DSC Normalization: 9b163  
 Sample Weight: 1.810 mg  
 Thu Jun 13 18:05:48 1991  
 TNT, SS Pans with O-rings  
 (Normalized)

PERKIN-ELMER  
 7 Series Thermal Analysis System  
 (L) indicates left ordinate; (R) indicates right ordinate  
 All runs at 235 C unless otherwise noted

DSC Normalization: 9b207  
 Sample Weight: 1.820 mg  
 Fri Jul 19 15:54:17 1991  
 NTO, SS Pans with O-rings  
 (Normalized)



910613-10, 1 sec at 235 C  
 Temp 1: 235.0 C Time 1: 50.0 min

Time (minutes)  
 Temp 1: 235.0 C Time 1: 50.0 min

Figure 6

## TERNARY PHASE DIAGRAM OF MEIC/NaCl/AlCl<sub>3</sub>

Dr Do Ren Chang  
Chemistry Department, Averett College, Danville, VA 24541

### Abstract:

Phase diagram of MEIC/NaCl/AlCl<sub>3</sub> ternary system for solid-liquid transitions has been investigated extensively using DSC. A contour map of isotherms has been constructed from these measurements. It reveals that the vast domain of this system is liquid at room temperature and along 0.5 mole fraction of AlCl<sub>3</sub> line parallel to the MEIC/NaCl binary is the ridge of the system. The MEIC/NaCl binary phase diagram has been studied also for the completeness. A new compound is formed with MEIC/NaCl in 2:1 ratio.

### Introduction:

Molten salts as electrolytes have been investigated extensively for battery applications [1]. Some molten salts possess a wide electrochemical window which may lead to a battery with high potential and high theoretical specific energy, thus it may be operated at high current densities without too high power losses because of the high conductivity of molten salts. However, most of the available molten salts batteries are operated at significant high temperature due to the electrolyte used. A molten salt system consists of MEIC (1-methyl-3-ethylimidazolium chloride) and AlCl<sub>3</sub> remaining in liquid at room temperature was first predicted and synthesized at FJSRL [2, 3]. These chloroaluminate melts not only have a wide electrochemical window but also are liquid at room temperature between  $N=0.3$  to  $0.7$ , where  $N$  is the mole fraction of AlCl<sub>3</sub> [4]. For  $N<0.5$  are called basic melts.



The main ionic species in this region are  $\text{MEI}^+$ ,  $\text{Cl}^-$  and  $\text{AlCl}_4^-$ ; and  $N > 0.5$  are called acidic melts, besides  $\text{MEI}^+$ , the dominating anion species are  $\text{AlCl}_4^-$  and  $\text{Al}_2\text{Cl}_7^-$ . The  $\text{Al}_3\text{Cl}_{10}^-$  appear when  $N > 0.75$  [5]. For  $N = 0.5$  is called a neutral melt and has been shown experimentally [6] to have the widest electrochemical window of approximately of 4.4 V. However, once an electrochemical reaction takes place in the cell, the electrolyte moves away from neutrality. Recently, NaCl was discovered to have buffer action in the neutral melts [7]. Relatively little is known about this ternary system of MEIC/NaCl/ $\text{AlCl}_3$ . This report details the phase behavior of this system in solid-liquid transition region. Along with the binary MEIC/NaCl presented here, together MEIC/ $\text{AlCl}_3$  [4] and NaCl/ $\text{AlCl}_3$  [8, 9] gave a complete picture of the ternary.

#### Experiment:

(1) Melts preparation: MEIC and aluminum chloride are prepared and purified as described in reference 2. Sodium chloride with 99.999% purity is purchased from Aldrich and is used directly without any treatment. All mixtures are prepared and weighed inside the glovebox under helium.

Appropriate different amounts of MEIC and NaCl are mixed first, then brought to melt under  $100^\circ\text{C}$ , cooled back to solid and then grinded into powder. These binary compositions are then used to prepare ternary compositions by adding different amounts of  $\text{AlCl}_3$ . Six binary series are selected in this study. They are A, C, E, G, I and K. corresponding to mole fraction of NaCl in MEIC/NaCl binary of about 0.1, 0.3, 0.4, 0.5, 0.6 and 0.7 respectively. When  $\text{AlCl}_3$  is added to each series, some form gel, some form clear liquid and some have undissolved excess  $\text{AlCl}_3$  at room temperature. Small heat is applied (less than  $50^\circ$  in temperature) to

help dissolve and mix. Since  $\text{AlCl}_3$  is highly volatile, temperature above  $60^\circ\text{C}$  is undesirable to avoid the change of composition in the sample. Some other binary  $\text{NaCl}/\text{AlCl}_3$  have also been used to prepare the ternary system in this study.

## (2) Measurement:

For visual observation, samples are loaded into the capillary tube, the open end of the tube is then temporarily sealed with paraffin film, then moved outside the glovebox and quickly sealed by flame. Melting point is observed with Meltemp device if the melting point is greater than the room temperature. For DSC measurement, about 10 mg samples are sealed in a steel pan fitted with an o-ring cover pan. Perkin-Elmer 7 series is used. This includes DSC-7, TAC-7 Instrument Controller, Perkin-Elmer 7500 professional computer, Perkin-Elmer graphics plotter and an intracooler with the capability to lower the temperature to  $-70^\circ\text{C}$ . Samples are scanned at a fixed rate of  $5^\circ/\text{min}$ . All binary samples showed an endo peak in the DSC run. The results are shown in Table 1 and Figure 1. A typical DSC run for a ternary sample can have one exo and several endo peaks, or one exo and one endo peak. If no peak appears this means that the melting point or glass transition temperature is very low, beyond the ability of the intracooler. In such a situation, the melting point is carried out with the following set up as shown in Figure 2. For viewing at subambient temperature, the outside air jacket needs to be sprayed with methanol occasionally. The results for the ternary are shown in Table 2.

## Result and Discussion:

For MEIC/ $\text{NaCl}$  binary, in general, the visual observation differs only slightly from DSC values. For a wide range of compositions, the melting points varied slightly.

This unexpected result indicates that the liquid melts are nearly pure MEIC, that is the very low solubility of NaCl. The existing local maximum at 33 mol % NaCl reveals a compound formation of 2 to 1 in MEIC to NaCl, with chemical formula  $(\text{MEI})_2\text{NaCl}_3$ . As shown in Figure 1, there exists an eutectic composition at 60 mol % NaCl and melts at 85°C. For ternary MEIC/NaCl/ $\text{AlCl}_3$  system, since experiments are carried out mostly in such a way from six preselected binary series, then by adding  $\text{AlCl}_3$  to each series. Thus, in the ternary phase diagram as shown in Figure 3, these paths correspond to by connecting points A, C, E, G, I and K on the MEIC/NaCl side to the apex  $\text{AlCl}_3$  respectively. Along a particular path, the ratio of MEIC to NaCl is fixed. The melting point will change according to different amounts of  $\text{AlCl}_3$  being added. The results are shown in Table 2. As an illustration for this G-series, the exo peak corresponds most likely to a glass to crystalline solid transition and the endo peaks correspond to the solid-solid phase transitions or to the solid-liquid transition if the endo peak occurs at the highest temperature. Based on these interpretations, the phase diagram along the G-series is constructed as shown in Figure 4. A local maximum is called peritectic point in ternary corresponds at the following compositions of mole fraction  $\text{AlCl}_3$  0.5, MEIC 0.25 and NaCl 0.25 melts at temperature 36°C. It is a new compound  $(\text{MEI})\text{Na}(\text{AlCl}_4)_2$  and a double salt of MEI  $\text{AlCl}_4$  and  $\text{NaAlCl}_4$ . There exists also another less pronounced peritectic point and several eutectic points of ternary. Most of them below the room temperature. If the peritectic point is approached from both sides of the G-line, the slope is rather steep. By combining all the data studied thus far, it showed that a maximum exists always at 0.5 mole fraction in  $\text{AlCl}_3$  in this ternary system. Along this 0.5 mole fraction  $\text{AlCl}_3$  line MN parallel to the MEIC/NaCl binary side in Figure 3, the ternary system can be regarded as a pseudo-binary of MEI  $\text{AlCl}_4$ /Na $\text{AlCl}_4$ . The melting point of this pseudo-binary

has been obtained [10]. The present result along the MN line is shown in Figure 5 for the comparison with the reference 10. The results confirm the general features of the pseudo-binary and this is a ridge line in the ternary system. As the peritectic point is approached from both sides of the MN line, the slope is much less steep than along the G-line. This indicates that a strong dependence on the  $\text{AlCl}_3$  component in the ternary system. This is understandable due to chemical reactions occurring between  $\text{AlCl}_3$  and MEIC and NaCl respectively. The agreement between the two results is remarkable.

The mapping of isotherms can be done in the ternary system by incorporating all three binaries involved. As shown in Figure 3, the contour lines of the melting point are made for  $20^\circ$  intervals for clarity in high temperature range and smooth interpolation can be made for intermediate temperature in low temperature domain. The shaded area shown in the figure has not been investigated due to excess insoluble components that exist at the preparation of the sample.

Conclusion: It reveals a new compound  $(\text{MEI})_2\text{NaCl}_3$  formed in binary MEIC/NaCl and a new compound  $(\text{MEI})\text{Na}(\text{AlCl}_4)_2$  in the ternary system. The vast domain of this ternary is in liquid form at room temperature. It is an excellent candidate as an electrolyte for molten salt battery system. For buffered neutral melts is concerned, a wide range of NaCl can be used.

Acknowledgment: The Air Force Office of Scientific Research is gratefully acknowledged for support of this work through RDL-SFRP program. Thanks to my focal point Dr John Wilkes for many discussions during the course of study. Finally, I would like to thank Fred Kibler for skillful glass blowing and Linda Pukajlo for her expert word processing of the manuscript.

#### References:

1. Reference in C.L. Hussey, in G. Mamantov and C.B. Mamantov eds. *Advances in Molten Salt Chemistry*, Elsevier, New York, 1983, Vol 5. p 185.
2. J.S. Wilkes, J.A. Levisky, R.A. Wilson and C.L. Hussey, *Inorg. Chem.* 21, 1263 (1982).
3. J.S. Wilkes, J.A. Levisky, J.S. Landers, C.L. Hussey, R.L. Vaughn, D.A. Floreani and D.J. Stech, FJSRL-TR-81-0011, Frank J. Seiler Research Laboratory, USAF Academy, October 1981.
4. A.A. Fannin, Jr., D.A. Floreani, L.A. King, J.S. Landers, B.J. Piersma, D.J. Stech, R.L. Vaughn, J.S. Wilkes and J.L. Williams, *J. Phys. Chem.*, 88 2614 (1984).
5. C.J. Dymek, Jr., C.L. Hussey, J.S. Wilkes and H.A. Oye, *Proc. Joint Int. Symp. on Molten Salts*, Mamantov et al eds. Vol 87-7. The Electrochemical Soc., 1987, pp 93-104.
6. M. Lipstajn and R.A. Osteryoung, *J. Electrochem. Soc.* 130, 1968 (1983).
7. T.J. Melton, J. Joyce, J.T. Maloy, J.A. Boon and J.S. Wilkes, *J. Electrochem. Soc.* 137, 3865 (1990).
8. R. Midorikawa, *J. Electrochem Soc., Japan*, 23 72 (1955).
9. V.I. Shvartman, *Zh. Fiz. Khim* 14, 253 (1940).
10. J.S. Wilkes, et al, unpublished results.

**Table 1**

Sample	Mole fraction of NaCl in the melt	Melting Point (°C)	
		Visual	DSC onset
A	0.1032	86.0	86.50
B	0.2021	85.8	86.49
C	0.3166	87.0	86.73
D	0.3526	86.6	87.06
E	0.3968	86.6	86.74
F	0.4601	86.8	86.59
G	0.4999	86.3	86.43
H	0.5515	86.0	85.88
I	0.6058	85.4	85.01
J	0.6520	86.0	85.79
K	0.7067	87.2	86.09
L	0.7541	87.3	86.37
M	0.8057	86.5	86.52
N	0.9020	87.0*	85.37*
O	0.9508	86.8*	84.88*
P	0.8557	86.6*	85.61*
MEIC	0	87.0	87.51

\*Samples N, O and P are not completely melted under 100°C during their preparation, but were ground to mix well. Under visual observation, these samples appear as wet at their respective temperatures. These are taken as a sign of a phase transition, even though the entire sample remains as solid upto 245°C and some decomposition is shown by brownish color.

# MEIC/NaCl Phase Diagram

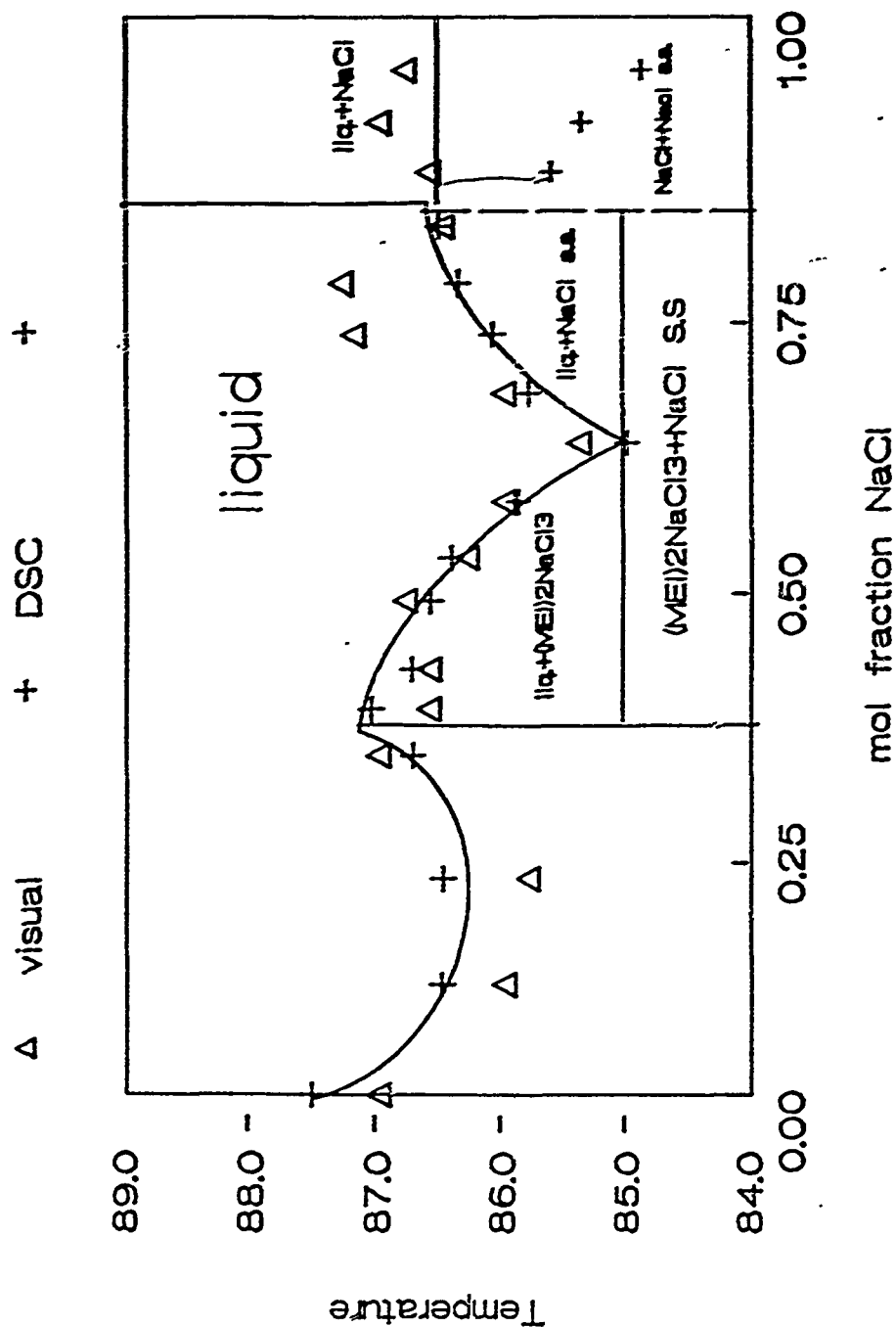


Figure 1. MEIC/NaCl binary phase diagram

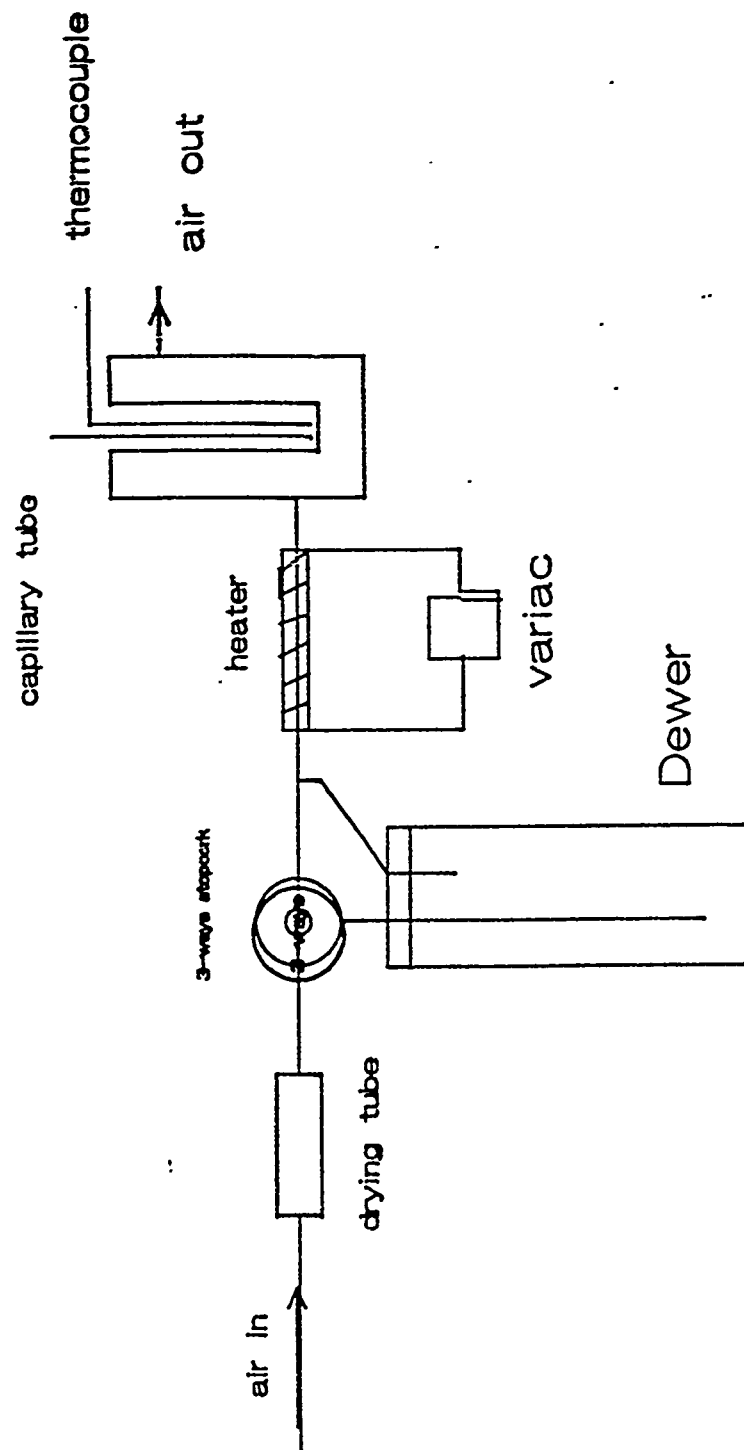


Figure 2. Melting point device



Figure 3. Ternary Phase Diagram

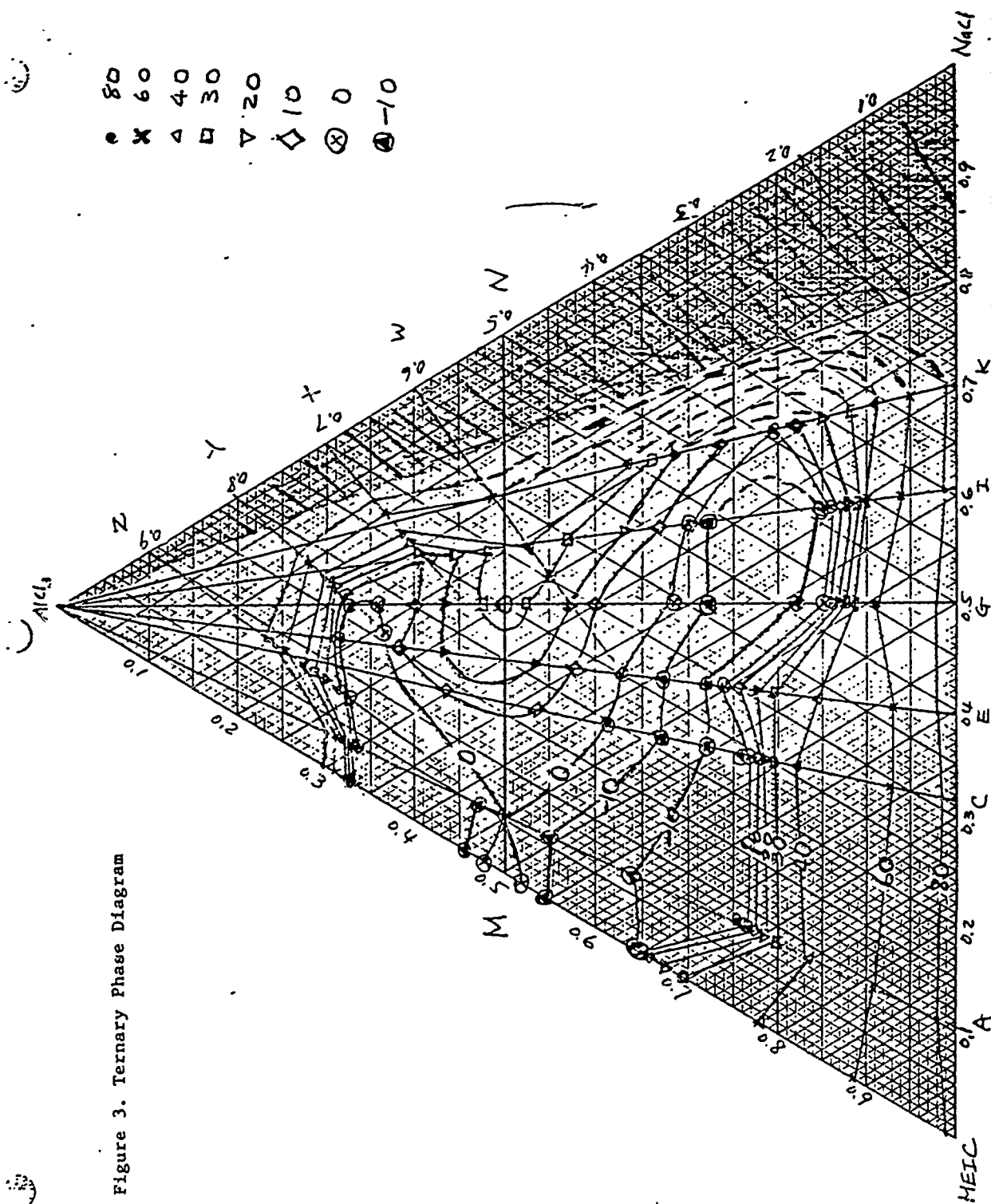
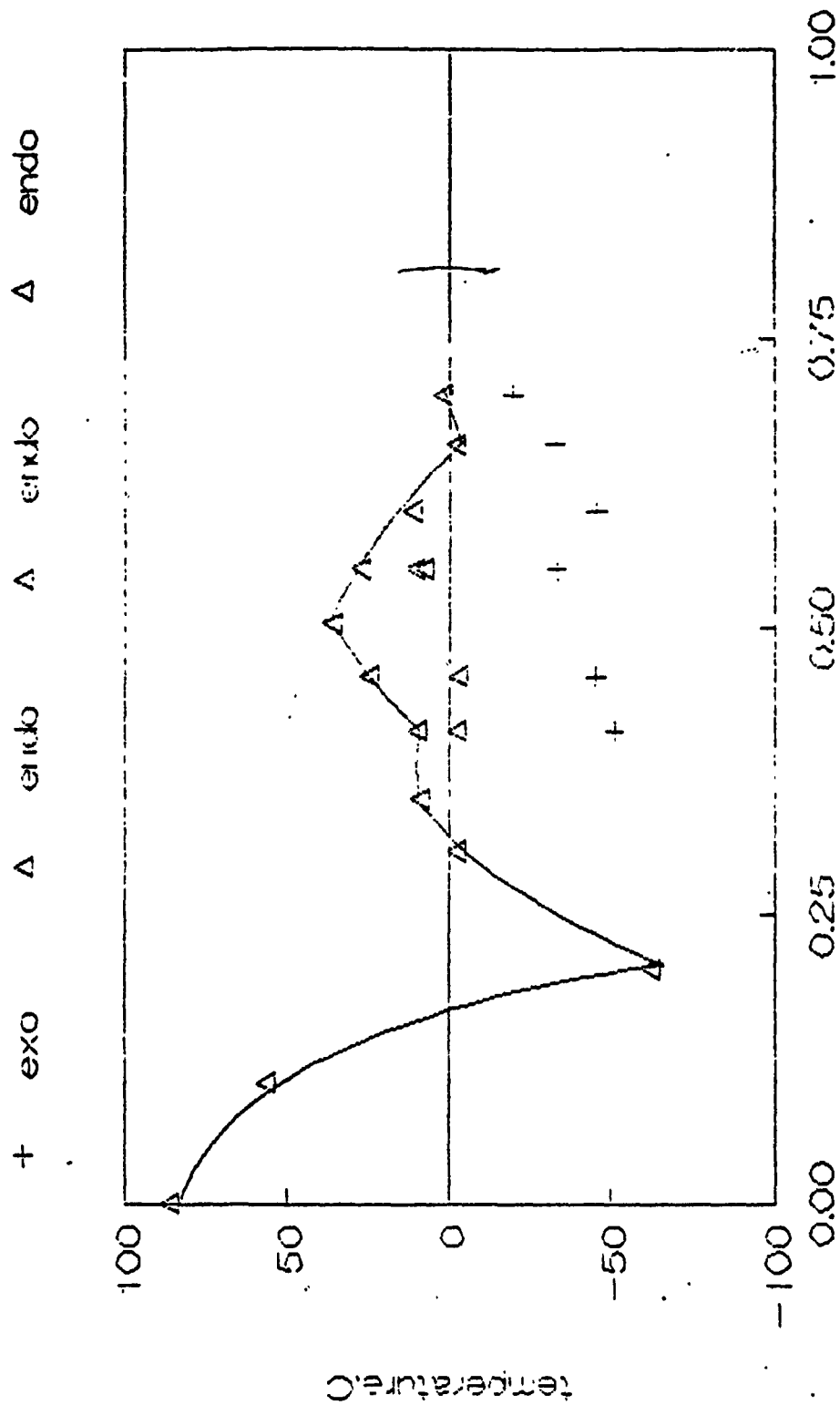


Table 2. Melting Points of the NEIC/NaCl/AlCl<sub>3</sub> Ternary

Mole Fraction AlCl <sub>3</sub>	A	C	E	G	I	K
0.10	4.543 <sup>a</sup> , 58.273	8.652 <sup>a</sup> , 52.139	25.735 <sup>a</sup> , 49.120	56.485	-0.620 <sup>a</sup> , 34.810, 41.982	-8.859 <sup>a</sup> , 37.509
0.20	9.553 <sup>a</sup> , 32.448	-1.797 <sup>a</sup> , 38.706	-4.378 <sup>a</sup> , 34.870	-62.5 <sup>b</sup>	-77.5 <sup>b</sup>	1.501
0.30	-61.5 <sup>b</sup>	-71.0 <sup>b</sup>	-43.536 <sup>a</sup> , -17.836	-2.320	-26.428 <sup>a</sup> , -1.373, 1.428	-29.399 <sup>a</sup> , -0.649 13.335
0.35		-14.029		9.236		
0.40	-45.786 <sup>a</sup> , -21.006	3.551	-21.273 <sup>a</sup> , -1.005 3.156	-51.210 <sup>a</sup> , -2.344, 10.198	-1.915, 28,343	-3.291, 38.545, 53.625
0.45				-45.490 <sup>a</sup> , -2.676, 25.151		
0.50	-0.413 (-0.1 <sup>b</sup> )	-49.636 <sup>a</sup> , -0.871, 14.153	-11.111 <sup>a</sup> , 27.247	36.448 (36.7 <sup>b</sup> )	34.115 (34.5 <sup>b</sup> )	33.503, 46.923, 55.082
0.55	-60.748 <sup>a</sup> , -15.217	-45.069 <sup>a</sup> , 12.955		-33.520 <sup>a</sup> , 8.192, 10.760, 27.885		
0.60	-94 <sup>b</sup>	-42.837 <sup>a</sup> , -2.689, 4.286	-40.021 <sup>a</sup> , 4.698, 16.794	-45.668 <sup>a</sup> , 11.506	-36.948 <sup>a</sup> , 3,088	
0.65	-94 <sup>b</sup>	-44.683 <sup>a</sup> , 4.772	-39.858 <sup>a</sup> , 1.694	-33.076 <sup>a</sup> , -2.043		
0.70	87.5 <sup>b</sup>	-49.066 <sup>a</sup> , 5.730	-39.343 <sup>a</sup> , -0.441	-20.0 <sup>a</sup> , 1.964	-39.714 <sup>a</sup> , 5.389	
a: exo peak, involving glass to crystalline transition likely b: visual observation						

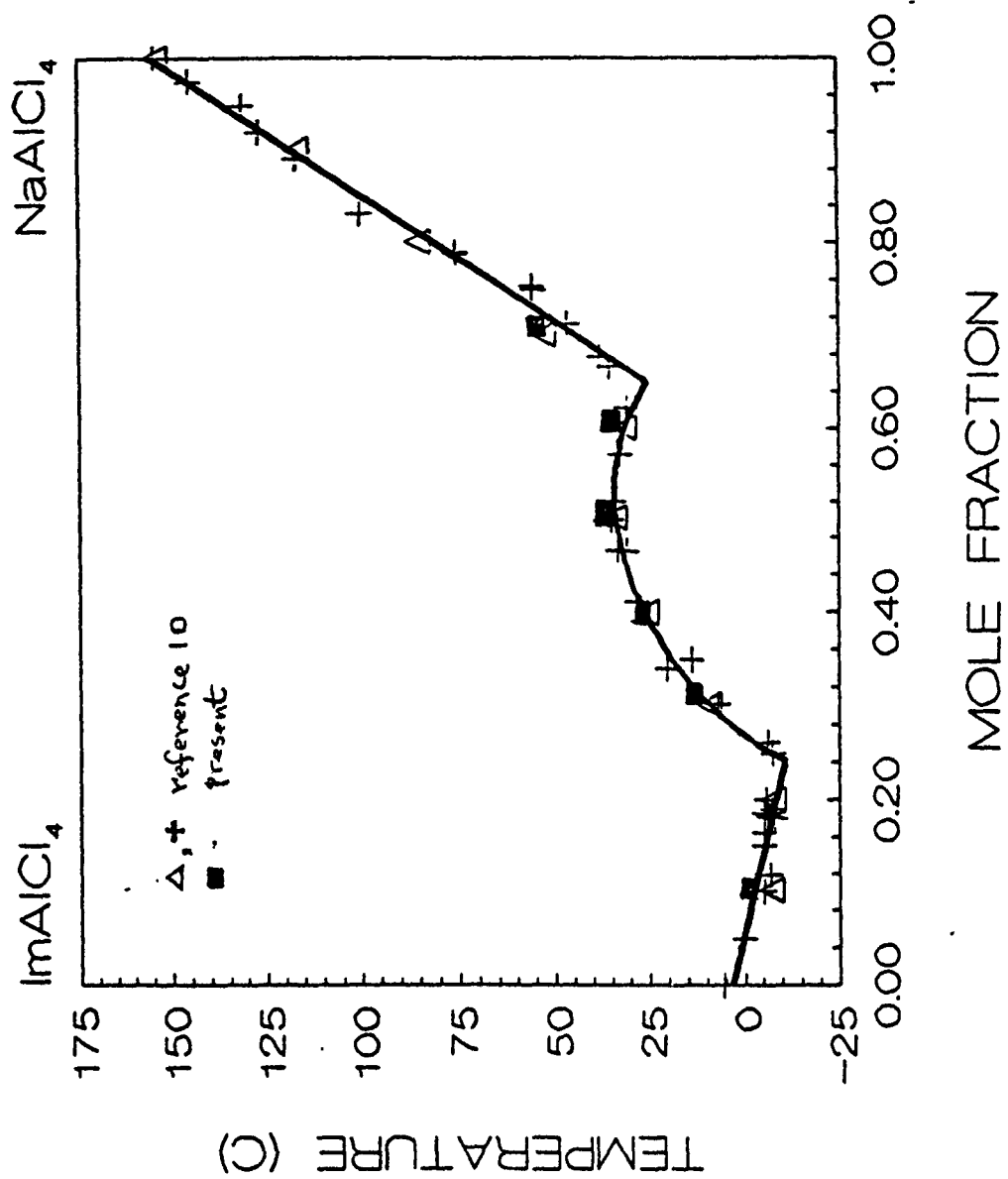
# meic/nacl/alcl3 ternary, G-series



mol fraction alcl3

Figure 4.

Figure 5.



# **PHOTO-ELECTRONIC NONLINEAR EXCITATIONS AND WAVE PROPAGATION IN PERIODICALLY MODULATED MEDIA**

**Dr. Marek Grabowski\***

*Frank J. Seiler Research Laboratory, US Air Force Academy, CO 80840-6528*

The propagation of electronic and electromagnetic waves in periodically modulated, nonlinear media is studied in terms of a discrete dynamical system. The analysis of this dynamical system is applied to the problems of tunneling of carriers in one-dimensional arrays of semimagnetic quantum dots, nonlinear photo-electronic excitations in saturable, dispersive and absorptive periodic medium, and bichromatic optical wave propagation in electrically poled waveguides. New results include the intensity dependent transmission coefficient, photonic bands with intensity dependent stop gaps, coherent soliton like states within forbidden states of linear theory, and phase-matched second harmonic generation in inversion symmetric systems.

---

\* permanent address: *Department of Physics, University of Colorado,  
Colorado Springs, CO 80933*

## I. INTRODUCTION

There are many very interesting phenomena associated with the propagation of electromagnetic and electronic waves in spatially periodic, nonlinear media. Recently, a number of these phenomena have been studied. In particular, Delyon *et al.*<sup>1</sup> have shown that a periodic modulation of a one-dimensional nonlinear medium induces multistability in the transmitted wave intensity, while Mills and Trullinger<sup>2</sup> have examined the localized, solitonlike structures associated with these systems. We have also studied these phenomena from the point of view of applications to the ballistic transport of electrons in semiconductor superlattices<sup>3</sup>, nonlinear transmission of light through multiple quantum well systems<sup>4</sup>, and universal dynamics of a discrete Hamiltonian map describing these systems<sup>5</sup>. Here we shall concentrate our attention on the photo-electronic excitations and intensity-dependent electromagnetic wave propagation in the presence of externally imposed spatial periodicity of the medium.

Specifically, we first investigate the transmission of waves via tunneling of carriers in quasi-one-dimensional arrays of semimagnetic quantum dots immersed in a nonmagnetic material. This quantum wire system is modeled by the appropriately derived Schrodinger equation, which is then transformed into a nonlinear complex map relating wavefunctions of the electrons on successive layers. The analysis of this map is then performed according to the methods previously described<sup>5</sup>.

Subsequently, we study the tunneling problem for the periodically modulated wire. We find that the large magnetopolaron (free carriers dressed by the induced polarization of the magnetic ions) effects due to the exchange interaction with magnetic ions result in a transmission via coherent, solitonlike electronic states. The calculated transmission coefficient shows the opening of intensity dependent gaps in the transmission spectrum. The experimental tunneling studies of, for example, CdMnTe quantum wires should demonstrate the here predicted effects of multistability, hysteresis, and the opening of stop gaps in the energy spectrum.

The research relevant to this part is described in more detail in the paper entitled *"Tunneling in a Periodic Array of Semimagnetic Quantum Dots"* which has been submitted for publication in Physical Review Letters and its manuscript is included in this report as the Appendix I.

It should be noted here that although this paper deals specifically with semimagnetic electronic systems, the problems investigated there are more general in nature and are relevant to a wide variety of systems, including light propagation in optical media with periodically varying dielectric constant. The effort in the direction of these kinds of applications is presently underway.

The studies so far were based on two basic assumptions: the modulation of the dielectric medium is on the scale of the wavelength of the propagating wave which is typically much greater than the length scale associated with both individual atoms and their separation, and the propagating wave is strictly monochromatic. Therefore, a classical description of the medium in terms of its polarizability entering a single wave equation was sufficient<sup>6</sup>. However, if one considers the interaction of individual atoms with the field, the quantum mechanical description of this interaction is necessary<sup>7</sup>. Consequently, in the second part of this research project we make the first step towards lifting the first of the above assumptions.

The first principle study of nonlinear photo-electronic excitations (polaritons) in saturable, dispersive and absorptive medium necessitates a self-consistent treatment of the propagating field and the quantum medium. The so-called Maxwell-Bloch equations<sup>7</sup> provide a full description of the spatial and temporal behavior of this coupled system. We investigate the transmission of photons through a one-dimensional array of polarizable quantum dots using these equations. This approach allows the study of transient effects which are crucial for optical pulses shorter than the relaxation times of the medium. Here, however, we have left the temporal behavior in the coherent regime to a subsequent projects and have concentrated on spatial behavior of relatively long pulses.

The Maxwell-Bloch equations for the quantum wire are solved in the steady state regime by a suitable reduction to a three-dimensional discrete map<sup>5</sup>. This nonlinear map admits stable linearized solutions in the form of polariton bands as well as nonlinear, coherent solutions corresponding to solitons and solitonic trains. The transmission properties of the system are studied via the threshold transmission coefficient and they exhibit transmitting solitonlike states in the polariton gaps as well as stop gaps in the polariton spectrum. These solitons and polaritons can be manipulated by separately contacting individual dots giving rise to a very exciting possibility of a tunable optical system.

More details of this research project are given in the paper entitled "*Nonlinear Polariton Excitations in Quantum Dot Arrays*", submitted for publication in Physical Review Letters, and enclosed here as the Appendix II.

Again, let us remark that the interesting results contained in this part of research are not limited to semiconductor quantum wires, but are also applicable to the problems of waves propagation in optical media. The future research in this area shall focus on these applications and should be extended into the full spatio-temporal regime of transient effects relevant to the propagation of short optical pulses.

Finally, the last project which, although not yet completed at the present time, will be continued in the future, dealt with the attempt at lifting the second approximation mentioned above. The simultaneous propagation of two waves: the fundamental and its second harmonic, was investigated in the presence of an external, periodic dc-field. This situation is particularly relevant to the problem of efficient second harmonic generation in the inversion symmetric media such as optical fibers<sup>8</sup> or possibly thin dielectric films acting as optical waveguides and periodically poled by a static electric field. The research in this area is being continued.



## References

1. F. Delyon, Yves-Emmanuel Levy, and B. Souillard, *Phys. Rev. Lett.* **57**, 2010 (1986).
2. D.L. Mills and S.E. Trullinger, *Phys. Rev. B* **36**, 947 (1987).
3. P. Hawrylak, M. Grabowski, and P. Wilson, *Phys. Rev. B* **40**, 6398 (1989).
4. Pawel Hawrylak and Marek Grabowski, *Phys. Rev. B* **40**, 8013 (1989).
5. Marek Grabowski and Pawel Hawrylak, *Phys. Rev. B* **41**, 5783 (1990).
6. Y.R. Shen, *Principles of Nonlinear Optics* (John Wiley & Sons, 1984).
7. Paul N. Butcher and David Cotter, *The Elements of Nonlinear Optics* (Cambridge University Press 1990).
8. Govind P. Agrawal, *Nonlinear Fiber Optics* (Academic Press Inc. 1989).

**TUNNELING IN A PERIODIC  
ARRAY OF SEMIMAGNETIC QUANTUM DOTS**

Pawel Hawrylak

Institute for Microstructural Sciences, National Research Council, Ottawa, K1A 0R6

Marek Grabowski

Department of Physics, University of Colorado, Colorado Springs, CO 80933

J.J. Quinn

Department of Physics, University of Tennessee, Knoxville, TN 37996

abstract

We investigate the tunneling of carriers in a quasi one dimensional array of semimagnetic quantum dots. The large magnetopolaron effects due to the exchange interaction of carriers with magnetic ions result in a transmission via soliton-like electronic states. The intensity dependent transmission coefficient shows the opening of intensity dependent gaps in the transmission spectrum.

## INTRODUCTION.

Magnetopolaron effects in the bulk semimagnetic semiconductors such as  $\text{Cd}_{1-y}\text{Mn}_y\text{Te}$  and quantum wells are well documented<sup>[1]</sup>. Magnetopolarons are free carriers dressed by the induced magnetic polarization field of the magnetic ions. Polaronic effects are weak in three dimensions and can only be observed with carriers bound to impurities. In quantum wells their stability is marginal<sup>(2)</sup>. In one dimensional systems these polaronic effects should be strong and lead to localized, soliton-like states<sup>[3]</sup>. The strength of polaronic effects should increase dramatically in quantum dots. We consider here an array of coupled quantum dots as a model of a periodic nonlinear system. Such system can be realized in a quantum wire with strongly varying concentration of magnetic ions.

## THE MODEL

Let us consider a  $\text{Cd}_{1-y}\text{Mn}_y\text{Te}$  quantum wire with an effective radius  $r_0$  normal to the growth direction the  $z$ -axis. The wire is built with unit cells of periodicity  $a$ . The unit cell of this wire consists of a dot of width  $d$  with low Mn concentration ( $y \sim 0.1$ ), and a barrier of width  $b$  and a high Mn concentration ( $y \sim 0.7$ ). In the barrier the antiferromagnetic interaction between Mn ions dominates and we expect the barrier to be in a "spin glass" phase. The dot with a low Mn concentration is assumed to be in a paramagnetic phase. In the "mean field" approximation, the effective Hamiltonian for the dot can be written as<sup>[4]</sup>:

$$H = \frac{p^2}{2m} - V_0 - V_m B_{5/2} \left( \frac{g_{\text{Mn}} \mu_B B_{\text{eff}}}{k_B T_{\text{eff}}} \right) \quad (1)$$

The first term is the kinetic energy,  $V_0$  is the barrier height (energy being measured from the top of the barrier),  $V_m = \frac{5}{2} \chi \beta N_0 J_z$  is the magnetic potential with  $\beta N_0$  being the exchange energy ( $\sim 880$  meV for the valence band)  $\chi$  is the average effective concentration

of Mn ions and  $J_z = \frac{3}{2}$  is the spin of free carriers (holes). The Brillouin function  $B_{5/2}$  describes the paramagnetic susceptibility of the Mn ion with spin  $5/2$  in the effective magnetic field  $B_{\text{eff}}$  at the effective temperature  $T_{\text{eff}}$ . The effective field  $B_{\text{eff}}$  acting on the ion is a sum of the external field  $B_0$  and the exchange field  $B_{\text{ex}} = \frac{5}{2} \beta N_0 J_z |\psi(z)|^2 / N_0$  which depends on the probability  $|\psi(z)|^2$  of the carrier being in the position of the magnetic ion.  $N_0$  is the number of cations per unit volume. The wavefunction for the carrier in the wire

is taken in the form  $\psi(r, z) = \frac{1}{\sqrt{\pi r_0^2}} e^{-(r/r_0)^2} \phi(z)$ . Substituting the wave function  $\psi$  into

Eq. 1, and retaining only linear terms in the expansion  $B_{5/2}(x) \sim .477x$ , gives a simple nonlinear Schrodinger equation for  $\phi(z)$  which can be written

$$-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial z^2} \phi(z) - \left\{ V_0 + V_B + \left( \frac{5}{4} \beta N_0 J_z / n_c \right) d |\phi(z)|^2 \right\} \phi(z) = E \phi(z) \quad \text{in the well,}$$

$$\text{and} \quad -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial z^2} \phi(z) = E \phi(z) \quad \text{in the barrier.} \quad (2)$$

Here  $n_c$  is the number of cations in the well,  $V_B = 0.467 \text{ g}_{\text{Mn}} \mu_B B_0 / k_B T_{\text{eff}}$  is the potential due to the external magnetic field while the last term is the self-consistent potential due to the exchange field. In the following we shall restrict ourselves to zero external magnetic field and narrow dots. To understand the combined effects of nonlinearity and periodicity it is sufficient to replace the effective potential of the well by a suitable chosen delta function potential. The resulting Schrodinger equation for a wire with  $N$  dots now takes the final form:

$$-\frac{\partial^2}{\partial z^2} \phi(z) - \left\{ \sum_{\ell=1}^N \left[ V_d + \alpha d |\phi(z)|^2 \right] \delta(z - \ell a) \right\} \phi(z) = E \phi(z) \quad (3).$$

The energy in Eq.3 is measured in the effective Rydbergs Ry and lengths in effective Bohr radii  $a_0$ . For holes in CdTe with an effective mass of 0.5 of the electron mass, the effective Rydberg  $Ry = 72 \text{ meV}$  and the Bohr radius  $a_0 = 10 \text{ \AA}$ . In Eq. 3 the function  $\phi$  is dimensionless ( $\phi \rightarrow \sqrt{a_0} \phi$ ), and the potential V has been chosen to reproduce the bound state of a single well of width d and barrier height  $V_0$ . For  $V_0=1Ry$ ,  $d=20\text{\AA}$ ,  $a=40\text{\AA}$ ,  $k_B T_{eff}=5\text{meV}$  we estimate  $V = 0.69 Ry$ . The coupling constant  $\alpha$  depends on the number of Mn ions, and for 10 cations per dot we estimate  $\alpha=10^{-2}Ry$ .

Eq.3 describes an array of coupled quantum dots. Because of the nonlinearity, the eigenstates of Eq.3 cannot be classified according to the Bloch scheme. We proceed by integrating Eq.3 across the  $\ell$ -th singularity. Requiring the continuity of the wavefunction  $\phi(\ell-) = \phi(\ell+)$ , and writing the energy E of a particle as  $E=k^2$ , one obtains a nonlinear complex map relating wavefunctions on successive layers:

$$\phi_{\ell+1} + \phi_{\ell-1} = 2 \left\{ \cos(ka) - \frac{Vda \sin(ka)}{2ka} - \frac{\alpha da^2 \sin(ka)}{2ka} |\phi_{\ell}|^2 \right\} \quad (4)$$

In the absence of nonlinearity ( $\alpha = 0$ ) the wavefunctions are Bloch states  $= \exp(ik\ell a)$  characterized by a Bloch index  $\kappa$ . The standard band structure is then obtained from Eq.4 with a single band for negative energies and an infinite number of bands and gaps for positive energies (above the barrier). Such a classification is possible because we only need the relation between the phase of the wavefunction and its energy. In the nonlinear case we must retain the information about the phase and the amplitude. This problem is simplified due to the global gauge invariance of Eq.4, i.e. the conservation of the current

$$J_{\ell} = -\frac{i}{2} \left\{ \phi_{\ell}^* \phi_{\ell+1} - \phi_{\ell+1}^* \phi_{\ell} \right\}. \text{ Consequently, the four-dimensional map of Eq.4 can be}$$

reduced<sup>[5]</sup> to a two-dimensional area preserving map with the current J and the energy E

as parameters. The classification of the solutions of Eq.4 is hence reduced to the study of the fixed points of this nonlinear map, and their basins of stability. This has been discussed in detail in Ref. [5]. The main effects are illustrated by considering the tunneling of electrons along a wire.

## THE TUNNELING

Let us consider a situation in which the periodically modulated  $\text{Cd}_{1-x}\text{Mn}_x\text{Te}$  section of the wire with  $N$  dots is inserted into the  $\text{CdTe}$  wire. The electrons traveling in a wire will be transmitted with the probability  $|T|^2$  and reflected with the probability  $|R|^2$ . Those electrons which are transmitted have traveled through a self-consistent band structure due to the exchange interaction with magnetic ions. In our model without exchange interaction, tunneling can occur if the energy of the transmitted particle is within the allowed energy spectrum of the Bloch band irrespective of the amplitude of the wavefunction or the flux of electrons. This is no longer true in the nonlinear case.

Let us first consider a single dot centered around  $z=0$ . We are interested in a state trapped in the dot. The wavefunction for this state has a form  $\phi(z)=A\exp(-q|z|)$  with energy  $E=-q^2$ . It is easy to see from Eq.3 that the corresponding energy is given by  $E=-\{Vd/(2-\alpha ad)\}^2$ . Increasing the nonlinearity  $\alpha$  lowers the energy. The first nontrivial case corresponds to two coupled dots. In this case there are two states. In Fig.1 we show the structure with two dots in the resonant tunneling geometry. The transmission problem is solved in the usual way: we assume that the wavefunction at the end of the wire has the form of the plane wave  $\phi(z) = T\exp(ikz)$ , where  $k$  is the wavevector corresponding to the energy  $E=-q^2$ , and  $T$  is the amplitude of the transmitted wave. We next iterate the wavefunction backwards, matching appropriately to the  $\text{Cd}_{1-x}\text{Mn}_x\text{Te}$  barriers. The inset (a) shows the transmission coefficient as a function of energy of incident particles in the

absence of nonlinearity. Two peaks correspond to two resonant states of a coupled dot system. In the presence of a nonlinear interaction the intensity of transmitted particles  $|\Gamma|^2$  becomes a multivalued function of the intensity  $|\Pi|^2$  of the incident particles, as shown in the inset (b), Fig.1. Hence the usual transmission coefficient  $|\Gamma|^2/|\Pi|^2$  depends on both the energy of the particle and transmitted intensity. When the number of dots  $N$  increases the energies form a band. The transmission coefficient becomes significant for the incident particle energy within the allowed energy band of the  $N$  dot array. In Fig.2 we show the transmission spectrum for a linear ( $\alpha=0$ , dashed line) and nonlinear ( $\alpha=0.01$ , solid line) array of 10 dots. The linear transmission shows an almost uniform spectrum of 10 peaks corresponding to the 10 possible states. The nonlinear spectrum is shifted to lower energies and shows only 3 distinct and well resolved peaks. This illustrates the opening of gaps i.e. forbidden, nontransmitting energy regions within linearly allowed band.

The qualitative features of the transmission problem are given by the "phase diagram" in parameter space of the energy  $E$  and amplitude  $|\Gamma|$  of the transmitted particle. To determine whether particle was transmitted, we iterate the wavefunction backward. The solutions which do not grow exponentially are considered to be transmitting. In this way we can map out the regions of parameter space with a finite transmission.

In Fig.3 we show such a diagram corresponding to particles incident on the array of 10 coupled dots for which transmission spectrum is shown in Fig.2. The dark areas correspond to transmitting states. We see clearly that the transmitting region breaks down into three distinct regions as shown in Fig.2. Hence the transmission via flux carrying states opens gaps in energy spectrum of the linear theory. The complex nature of this phase diagram is associated with the analysis<sup>(5)</sup> of the nonlinear dynamical map given by Eq.4.

In summary, we have shown that semimagnetic semiconductor quantum dots should exhibit interesting features due to strong magnetopolaronic effects. The tunneling studies should demonstrate the multistability, hysteresis, and the opening of gaps in the energy spectrum.



## References

- 1.J.K. Furdyna, J. Appl. Phys. 64, R29 (1988).
2. X.Zhu,J.J.Quinn, private communication
- 3.J. Spalek, Phys. Rev. B 30, 5345 (1984).
- 4.Ji-Wei Wu, A.V. Nurmikko, and J.J. Quinn, Solid State Comm. 57, 853 (1986),  
D. Heiman, P.A. Wolff, and J. Warnock, Phys. Rev. B27, 4848 (1983).
- 5.Marek Grabowski and Pawel Hawrylak, Phys. Rev. B41, 5783 (1990).

**Figure captions.**

Fig.1. The schematic picture of the energy diagram of two quantum dots ( dashed lines ) in a scattering geometry. The inset (a) shows the transmission coefficient  $|T|^2/|I|^2$  as a function of energy  $E$ . The inset (b) shows the incident intensity  $|I|^2$  as a function of transmitted intensity  $|T|^2$  in the nonlinear case (  $\alpha=.01$  ).

Fig.2. The transmission coefficient  $|T|^2/|I|^2$  as a function of energy  $E$  for 10 dots : linear spectrum - solid line; nonlinear spectrum (  $\alpha=.01$  )-dashed line.

Fig.3 The "phase diagram"  $\alpha a^2 |T|^2$  vs  $E$  of parameters corresponding to transmitted particle with energy  $E$  and amplitude  $|T|^2$  .

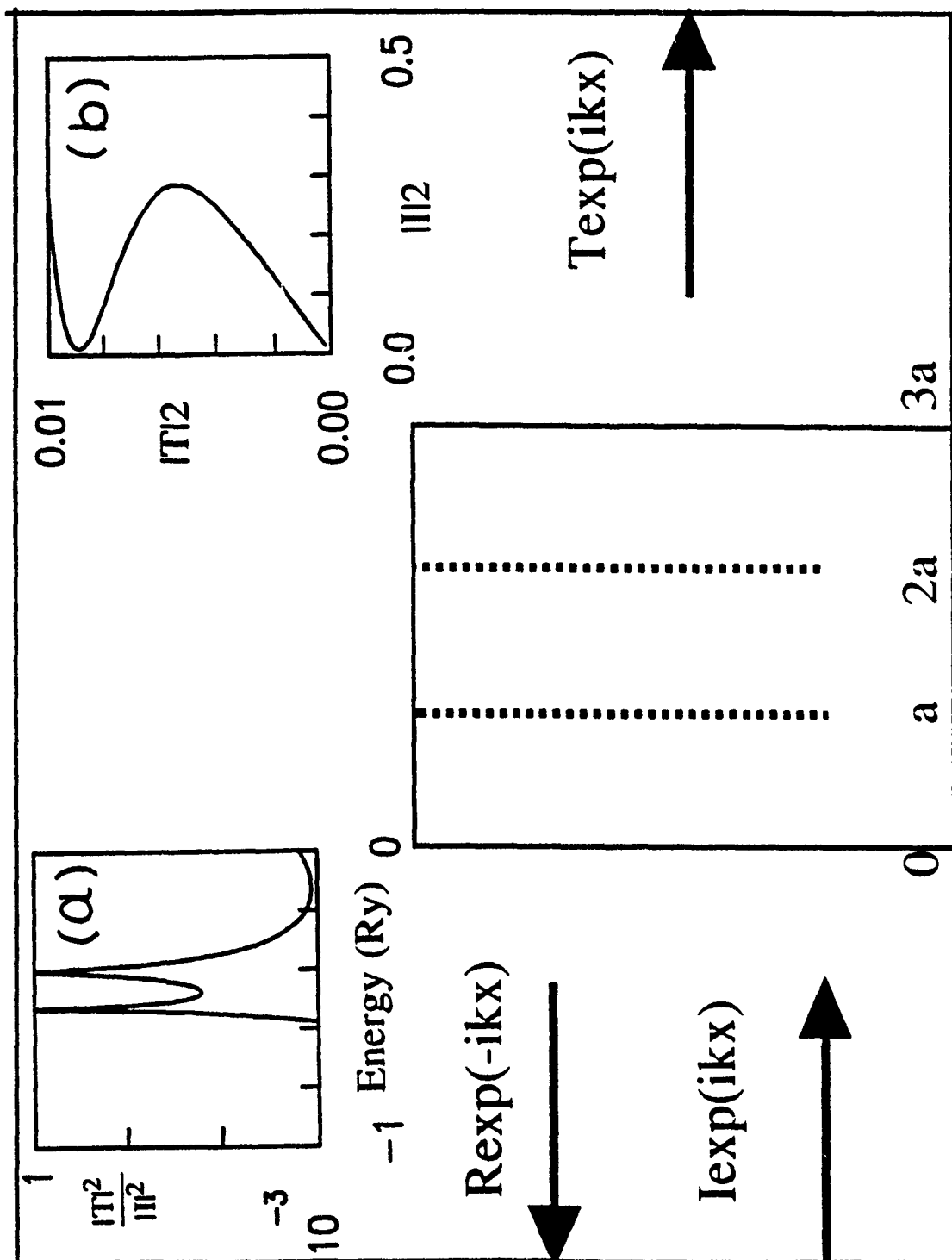
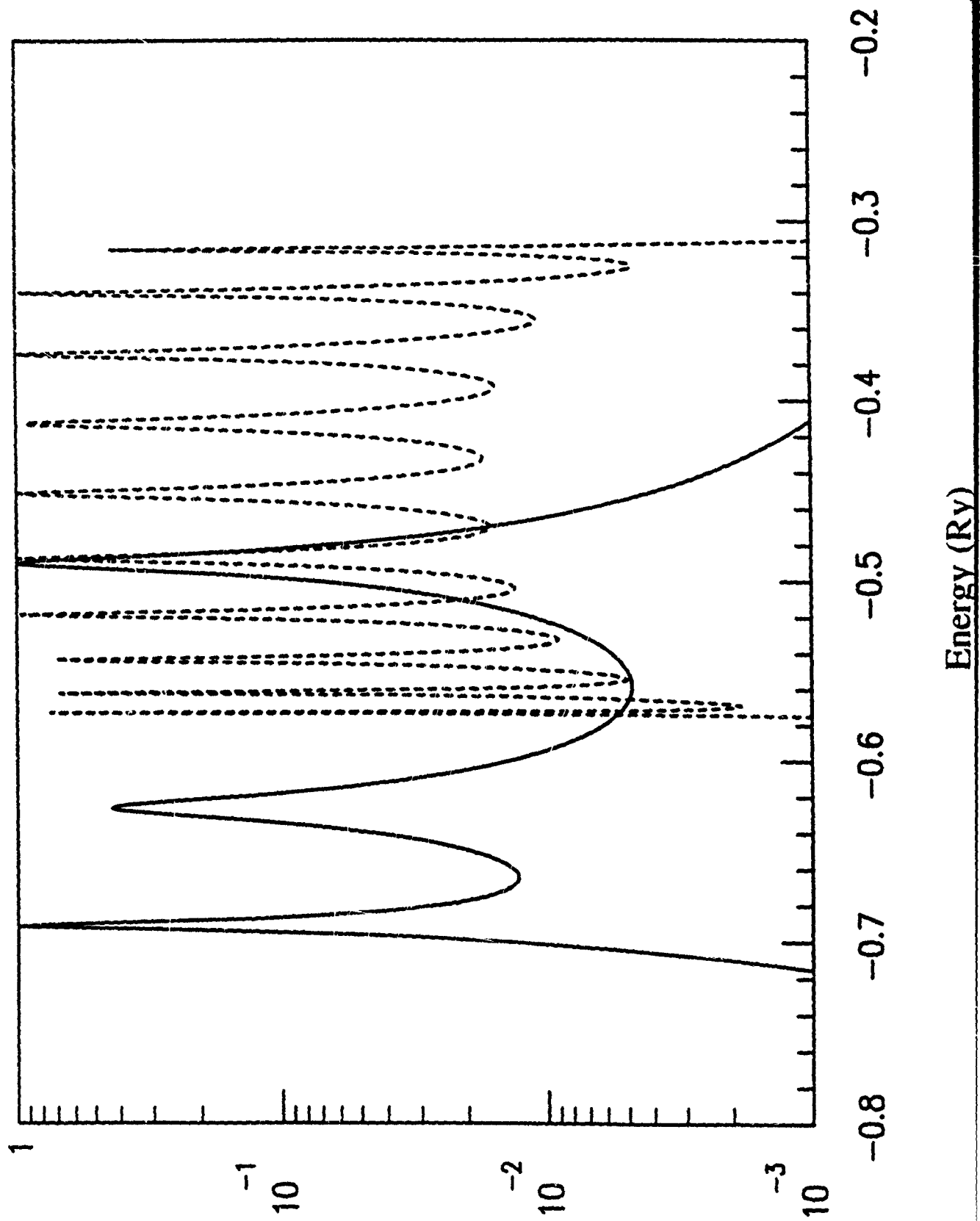
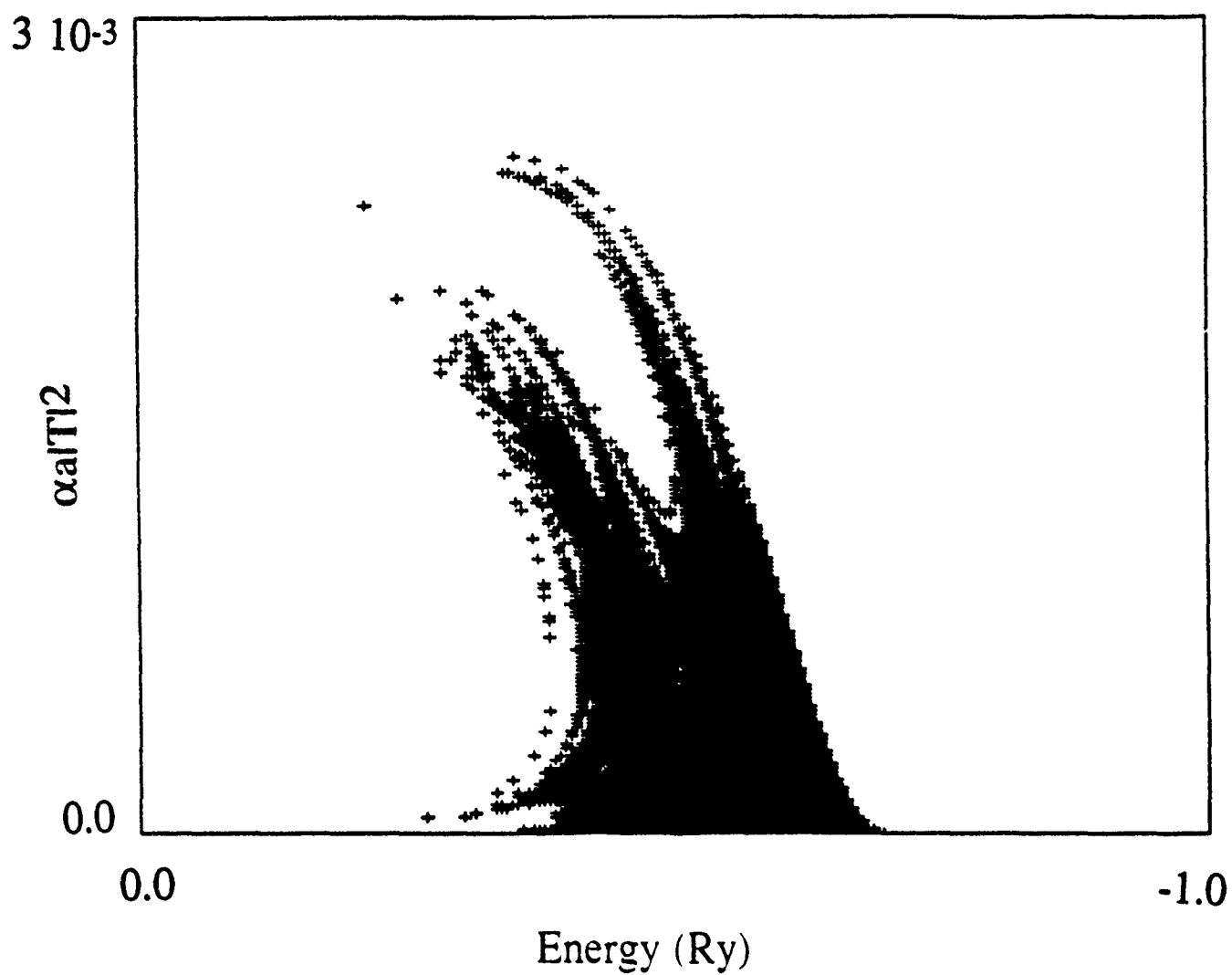


Fig. 1



TlII  
HII



# **NONLINEAR POLARITON EXCITATIONS IN QUANTUM DOT ARRAYS.**

**Pawel Hawrylak**

**Institute for Microstructural Sciences, National Research Council of Canada,**

**Ottawa, Canada, K1A 0R6.**

**Marek Grabowski**

**Frank J. Seiler Research Laboratory, US Air Force Academy,**

**Colorado Springs, CO 80840\*,USA.**

**Jacek A. Tuszynski**

**Department of Physics, University of Alberta,**

**Edmonton, Alberta, T6G 2J1, Canada.**

## **abstract**

**We investigate the tunneling of photons in a quasi one dimensional array of polarizable quantum dots using Maxwell-Bloch equations. This allows a first principle study of nonlinear polaritons in saturable, dispersive and absorptive, periodic medium. The nonlinear field equation for the steady state generates linearized excitations which form photonic bands and nonlinear excitations : gap solitons. The possibility of manipulating the photonic gaps and gap solitons by separately contacting the dots is examined.**

**\*Permanent address: Department of Physics, University of Colorado, Colorado Springs,  
CO 80933.**

The interaction of photons with electronic excitations in an artificially structured dielectric medium leads to many interesting and important phenomena; photonic gaps and gap solitons<sup>(1,2)</sup> providing good example. The modulation of the dielectric medium is on the order of the wavelength of light, typically much greater than the length scale associated with both individual atoms and their separation. Hence a classical description of the medium in terms of its polarizability is sufficient. If one considers the interaction of individual atoms with the field, the quantum mechanical description of this interaction leads to many interesting phenomena such as Rabi oscillations, photon echos, and self-induced transparency<sup>(3)</sup>. The dielectric medium and individual atom phenomena can be combined by creating arrays of mesoscopic atoms<sup>(4)</sup> ( quantum dots ) frozen in a semiconductor matrix. A typical separation  $a$  between dots can be comparable to the wavelength corresponding to the lowest radiative transition of the dot at frequency  $\omega_0$ . Hence one must describe self-consistently spatial and temporal collective behaviour of the field and quantum dots. This behaviour can be modified by selectively contacting individual dots and changing their equilibrium state by, for example, electrically injecting electron hole-pairs. To describe the essential features of such a complex system we shall adopt a model one dimensional array of two-level dots. Each quantum dot is characterized by its complex polarization  $P(x,t)$  and population inversion  $\delta n(x,t)$ . The time  $t$  is measured in units of characteristic transition frequency  $\omega_0^{-1}$  and distance  $x$  in units of the separation of dots  $a$ . The collective state of quantum dots and the linearly polarized complex electromagnetic field  $E(x,t)$  is described by the Maxwell-Bloch equations<sup>(3,5)</sup>:

$$\begin{aligned}
E_x(x,t) - \frac{1}{v^2} E_x(x,t) &= \frac{\beta^2(x)}{v^2} P_x(x,t) \\
\dot{P}_i &= -i(1-i\gamma)P + \frac{i}{2}(E+E^*)\delta n \\
\dot{\delta n}_i &= -i(E+E^*)(P^*-P) - \Gamma(\delta n - \delta n^0)
\end{aligned} \tag{1}$$

Here  $v$  is the effective velocity of light,  $\gamma$  and  $\Gamma$  are the inverse of relaxation rates of polarization and population inversion, respectively, and  $\delta n^0$  is the equilibrium population inversion in the absence of the field.  $\delta n^0=1$  corresponds to the ground state being fully occupied while  $\delta n^0=-1$  corresponds to fully inverted state. Different  $\delta n^0$  might be obtained by e.g. separately contacting individual dots and electrically injecting electron-hole pairs. The polarization  $P(x,t)$  is coupled to the field  $E(x,t)$  via the spatially dependent coupling constant  $\beta$  (proportional to the dipole moment of the dot) of the form

$$\beta^2(x) = \beta^2 \sum_{l=1}^N \delta(x-l).$$

All parameters of the theory can be determined experimentally by studying the linear response of the system.

The Maxwell-Bloch equations provide a full description of the spatial and temporal behaviour of the field and the dots. In the spatially homogeneous medium (the Jaynes-Cummings model) the full solution of the Maxwell-Bloch equations exhibits chaotic behaviour<sup>(6)</sup>. The spatio-temporal behaviour for a pulse width shorter than the relaxation time in a homogeneous medium exhibits self-induced transparency<sup>(7)</sup>. Even in a homogeneous medium the nonlinear interactions lead to nontrivial selection rules for the solitary waves<sup>(8)</sup>. We shall leave the temporal behaviour in the coherent regime to a separate presentation and concentrate here on spatial behaviour of pulses long enough for transient effects to be unimportant. In the rotating wave approximation the stationary solution  $P(x)$  of the Bloch equation for oscillating fields  $E(x,t)=\exp(-i\omega t)E(x)$ ,  $P(x,t)=\exp(-i\omega t)P(x)$  relates the polarization  $P(x)$  to the local field  $E(x)$  via the field dependent susceptibility  $\kappa(E)$ :



$$\kappa(E) = \frac{1}{2} \frac{(1 - \omega) + i\gamma}{\{(1 - \omega)^2 + \gamma^2 + (\gamma / \Gamma) E E^*\}} \delta n^0 \quad (2).$$

This complex susceptibility reflects the transition from saturable dispersive to saturable absorptive nonlinearity as the frequency sweeps through the resonance in the unbiased dot. Note that depending on the value of equilibrium population inversion the susceptibility of the individual dot can change from absorption to gain. In the steady state the nonlinear field equation determines the spatial distribution of the field:

$$E_{xx}(x) + k^2 E(x) = -k^2 \beta^2(x) \kappa(E) E(x) \quad (3).$$

The wavevector  $k = \omega/v$  corresponds to the propagation of the field between the dots. It is convenient<sup>(9)</sup> to define the field degrees of freedom in terms of its amplitude  $A(x)$  and phase  $\phi(x)$  :  $E(x) = A(x) \exp(i\phi(x))$  , and to identify them with coordinates of a fictitious particle. The spatial coordinate can now be identified with time. The corresponding momenta are defined as  $p = A_x$  and  $j = A^2 \phi_x$  . The variable  $j$  is just the energy flux:

$j = -(i/2) \{E^* E_x - E E_x^*\}$  . The equation of motion for the field now takes a very simple and intuitive form:

$$\begin{aligned} p_x &= -k^2 A + \frac{j^2}{A^3} - k^2 \beta^2(x) \operatorname{Re}(\kappa(A)) A \\ j_x &= -k^2 \beta^2(x) \operatorname{Im}(\kappa(A)) A^2 \end{aligned} \quad (4).$$

The first equation describes a harmonic, isotropic oscillator in the presence of a time dependent nonlinear force proportional to the dispersive part of the susceptibility and in the presence of a time dependent "centrifugal force" ( the term proportional to  $j^2/A^3$  ). The variable  $j$  can be interpreted as the angular momentum of the oscillator. The second equation describes the evolution of the momentum  $j$  and is governed purely by the absorptive part of the susceptibility. Clearly from the form of the susceptibility ,Eq.2, and Eq.4 it is obvious that the losses lead to a collapse while gain leads to a growth of the centrifugal barrier of the harmonic oscillator. In the absence of losses ( gain )  $j = \text{const}$  and

the case of small amplitudes ( nonsaturable nonlinearity ) , Eq.4 has been studied in detail in Ref(9) . Following Ref(9) the solution of Eq.4 can be written in a form of a compact 3-dimensional nonlinear map:

$$\begin{aligned} R &= R(Q^2 + J^2 / R^2) \\ Q &= 2c - sk\beta^2 \kappa'(R) - \frac{R}{R'} Q \\ J &= J - ks\beta^2 \kappa''(R) R \end{aligned} \quad (5),$$

where  $R=A^2$  ,  $Q=sp/kA+c$  ,  $J=js/k$  ,  $s=\sin(k)$  ,  $c=\cos(k)$ , and  $\kappa'$  and  $\kappa''$  stand for the real and imaginary parts of the susceptibility. Eq.5 allows for the analysis of the global properties of the nonlinear wave equation.

Let us first consider the stability of a plane wave  $E(x)=T\exp(ikx)$ . Starting with the initial condition  $R=T^2$ ,  $Q=\cos(k)$ ,  $J=T^2\sin(k)$  we iterate the map given by Eq.5.  $N$  times. The orbits whose amplitude  $R$  remains within a fixed radius  $R_{\max}$  from the initial condition are classified as stable. Hence we can create a phase diagram  $(T^2, \omega)$  of stable orbits. Fig.1 we show a phase diagram corresponding to parameters  $N=11$ ,  $v=0.2$ ,  $\beta=1.1$ ,  $\gamma=\Gamma=0.1$ ,  $R_{\max}=10^2$ . All dots in the absence of the field  $E(x)$  are in their ground state. The black crosses in Fig.1 correspond to stable orbits. Fig.1. shows two wide photonic gaps on both sides of the resonant frequency. These gaps begin to fill up with stable states -gap solitons. The region in the vicinity of the resonant frequency is quite complex ( disordered ) for low amplitude of the plane wave. As the amplitude  $T^2$  increases the absorptive losses saturate and plane waves with larger amplitudes become more stable. Eq.5 and the phase diagram it generates can be used to study the transmission experiment. However , we shall adopt a more standard ( but entirely equivalent ) approach to the transmission problem in terms of transfer matrices.

We decompose the field  $E(x)$  between  $(n-1)$ -st and  $(n)$  dot in terms of forward ( $A_{n-1}$ ) and backward ( $B_{n-1}$ ) propagating components:  $E(x)=A_{n-1}\exp(ik(x-(n-1)))+B_{n-1}\exp(-ik(x-$

(a-1)). Following Ref(10) we cast the problem of finding the solutions to Eq.3 in terms of the transfer matrix M, defined by:

$$\begin{pmatrix} A_{n+1} \\ B_{n+1} \end{pmatrix} = \begin{pmatrix} (1 - ik \frac{\kappa(E_n)}{2} \beta^2) e^{-ikz} & -ik \frac{\kappa(E_n)}{2} \beta^2 e^{-ikz} \\ k \frac{\kappa(E_n)}{2} \beta^2 e^{ikz} & (1 + ik \frac{\kappa(E_n)}{2} \beta^2) e^{ikz} \end{pmatrix} \begin{pmatrix} A_n \\ B_n \end{pmatrix} \quad (6).$$

The linear optical properties are obtained from this transfer matrix using linearized susceptibility. The polariton Bloch wavevector Q can be easily computed from the trace of the transfer matrix M:  $2 \cos Q = \text{tr}(M)$ . The effect of quantum dots is to renormalize the photon group velocity at the bottom of the polariton band:  $Q = k \sqrt{1 + \beta^2/2}$ , and to introduce gaps in the polariton spectrum. The lowest polariton gap falls in the frequency range:  $v\pi\beta^2 / \{\beta^2 + 2(1 - v\pi)\} < \omega < v\pi$ . This allows to determine two parameters of the theory:  $v$  and  $\beta$ . For parameters corresponding to Fig.1 the lowest polariton band is in the frequency range (0.36,0.63).

To calculate the transmission coefficient we start with a plane wave solution at the end of the sample in the form of a transmitted wave  $T \exp(ikx)$ . This solution is iterated backwards and matched across the first layer to the wave composed of incident I and reflected R components:  $E(x) = I \exp(ikx) + R \exp(-ikx)$ . The transfer matrix across the first layer is slightly different from the transfer matrix M. In this way we can determine the incident wave intensity  $|I|^2$  as a function of transmitted intensity  $|T|^2$  for each frequency  $\omega$  and number of layers N. Several such functions for selected frequencies are shown in Fig.2. For frequencies in the photonic band of the linear theory the transmitted intensity is a linear function of the incident intensity. For higher incident intensity the transmitted intensity  $|T|^2$  becomes a multivalued function of the incident intensity  $|I|^2$  in a usual way. For frequencies in the gap of the linear theory the incident intensity  $|I|^2$  goes toward a constant as the transmitted intensity  $|T|^2$  vanishes. Hence a finite incident intensity is needed

for a finite transmission to occur. The photonic gaps can be penetrated by a beam of sufficient intensity by propagating a solitary wave.

The situation in Fig.2. resembles bifurcation phenomena in equilibrium phase transitions: the multivaluedness in the band of linear theory is reminiscent of first order phase transitions, and the multivaluedness in the gap is reminiscent of second order phase transitions. Naturally the behaviour of the system will be different for increasing or decreasing incident intensity. Even more complex behaviour is expected in the vicinity of the resonance ( see Fig.1. ). Is it then possible to define a meaningful transmission coefficient  $|T|^2/|I|^2$  as a function of frequency  $\omega$ ? We propose to characterise the system by a *threshold transmission coefficient*. The threshold transmission coefficient corresponds to the ratio of the transmitted intensity to the minimum incident intensity required for finite transmission. In short, for each frequency we construct the incident intensity as a function of the transmitted intensity and find a minimum. The threshold transmission coefficient corresponds to this minimum. We show the linear transmission coefficient (Fig.3a) and the threshold transmission coefficient (Fig.3b) as a function of frequency for the same set of parameters as in Fig.1 and Fig.2. The threshold transmission coefficient illustrates nicely the effects of nonlinearity in terms of filling up the polariton gaps with nonlinear, soliton like excitations. Fig.3c illustrates briefly the effect of altering the equilibrium value of inversion of every other dot from  $\delta n^0=1$  to  $\delta n^0=0$ . This essentially removes half of the dots and leads to changes in the transmission coefficient.

In summary, we studied Maxwell-Bloch equations for a periodic array of two level quantum dots. The spatial distribution of the electromagnetic field "dressed up" by induced polarization of quantum dots is described by a nonlinear wave equation. Global solutions of the wave equation are equivalent to the study of a three dimensional nonlinear map representing all irreducible variables. This nonlinear map admits stable linearized solutions

as polariton bands. The nonlinear solutions corresponding to periodic orbits of the map describe solitons. Energy carrying solitons lead to transmission in the polariton gaps and gaps in the polariton bands. Polaritons and solitons can be manipulated by separately contacting individual dots leading to a tunable optical system.

The Coulomb effects associated with biexcitons in individual quantum dots<sup>(11)</sup> neglected here can be incorporated into a more accurate susceptibility.

## References:

1. W.Chen and D.L.Mills , Phys.Rev.Lett.**58**,160(1987).
2. C.Martijn de Sterke and J.E.Sipe , Phys.Rev.A,**43**,2467(1991)
3. L.Allen, J.H.Eberly, Optical Resonances and Two-Level Atoms, John Wiley & Sons Inc, 1975.
4. M.A.Reed, R.T.Bate, K.Bradshaw, W.M.Duncan, W.R.Frensley, J.W.Lee. and H.D.Shih, J.Vac.Sci.Technol.**4**,358(1986).; K.Kash, A.Scherer, J.M.Worlock, H.G.Craighead, and M.C.Tamargo, Appl.Phys.Lett.**49**,1043(1986).
5. M.Lindberg and S.W.Koch, Phys.Rev.B**38**,3342(1988); P.Hawrylak, Proceedings of NATO Workshop on Light Scattering in Superlattices, Mt.Tremblant, Que., Plenum Publ., 1992.
6. P.W.Milonni , M.L.Shih, J.R.Ackerhalt , Chaos in Laser -Matter Interactions, World Scientific Lecture Notes in Physics, Vol.6, World scientific Publishing, 1987.
7. S.L.McCall and E.L.Hahn, Phys.Rev.**183**,457(1969).
8. Spiros Branis , Olover Martinet and Joseph L.Birman, Phys.Rev.A,**43**,1549(1991)
9. Marek Grabowski and Pawel Hawrylak, Phys.Rev.B**41**,5783(1990).
10. Pawel Hawrylak and Marek Grabowski, Phys.Rev.B**40**,8013(1989).
11. Y.Z.Hu, S.W.Koch and D.B.Tran Thoai, Mod.Phys.Lett.**4**,1009(1990).

**Figure captions:**

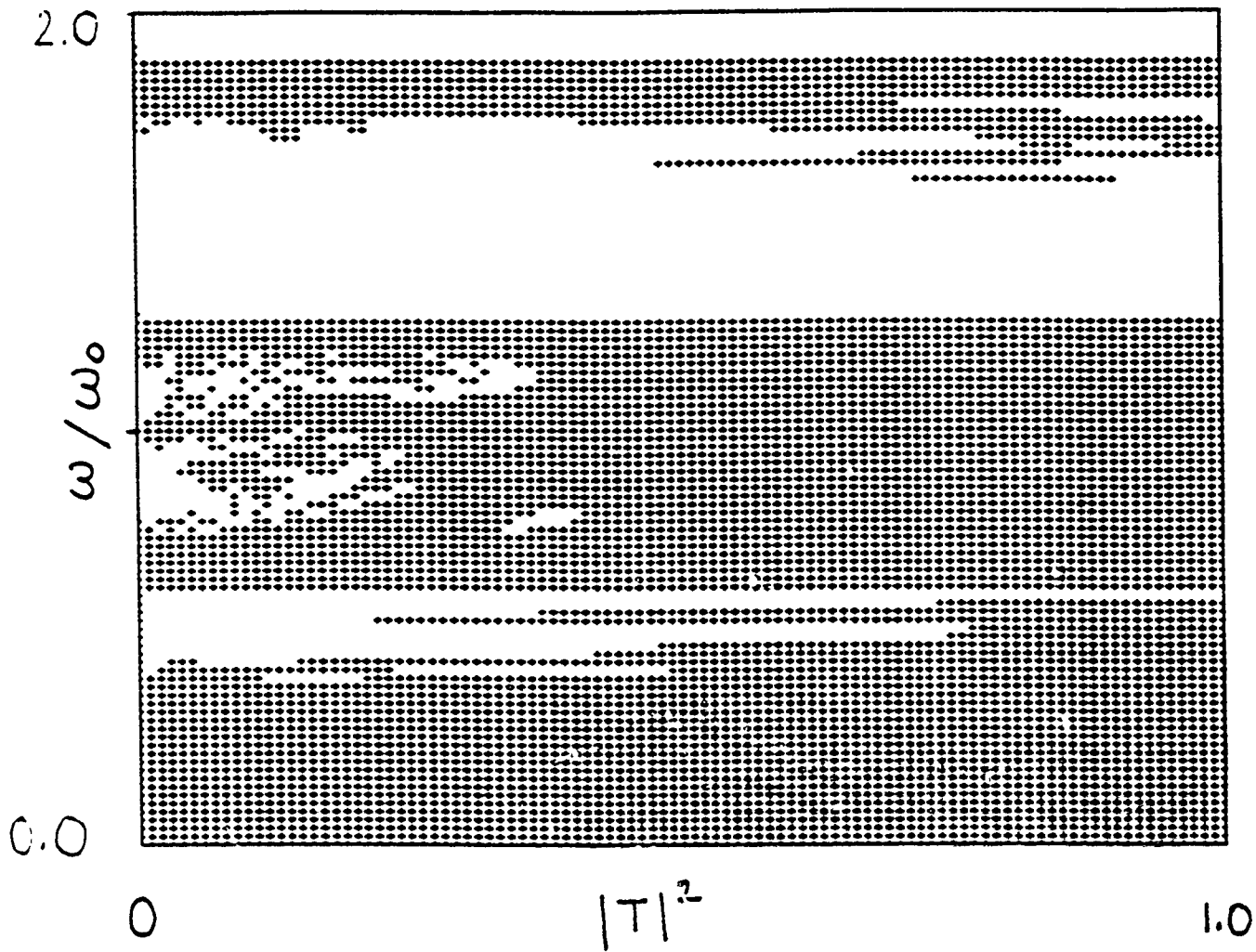
**Fig.1 Stability diagram ( $T^2\omega$ ) of the plane wave.**

**Fig.2. Incident intensity  $|I|^2$  vs transmitted intensity  $|T|^2$  for a set of frequencies**

**$\omega=.05,10,\dots,70$**

**Fig.3. Threshold transmission coefficient as a function of frequency  $\omega$  for (a) linear system**

**(b) nonlinear but unpolarized system (c) polarized nonlinear array.**



F. 9 1



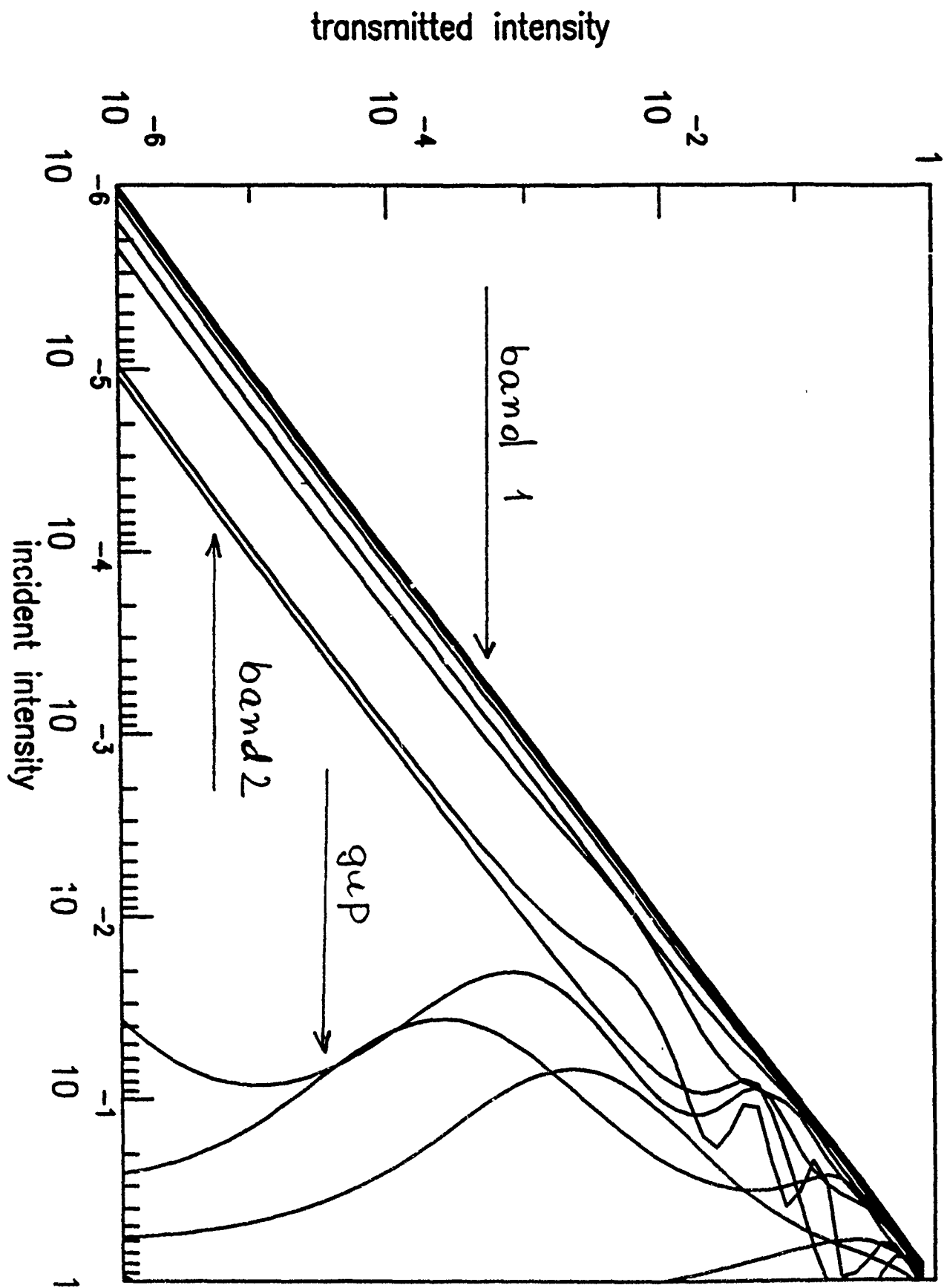
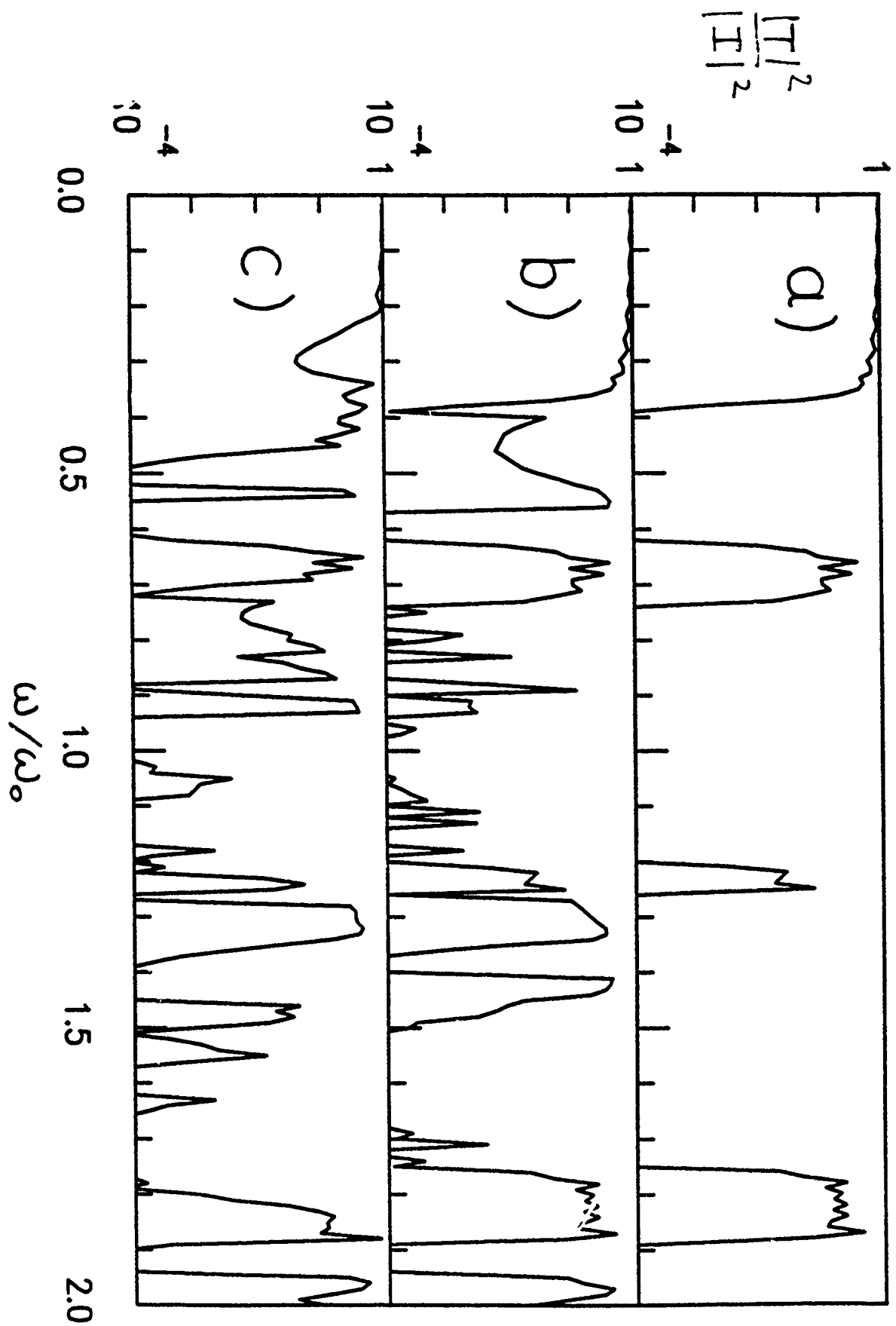


Fig.2.



# **1991 USAF Summer Research Program**

**Sponsored by the**

**Air Force Office of Scientific Research**

**Conducted by**

**RDL**

## **FINAL REPORT**

**An Ab Initio Study of the Adducts**

**between HF and HCl**

**and**

**Aluminum Hydrides, Halides, Hydroxides, and Oxides**

**Prepared by:** Dr. Gilbert J. Mains

**Academic Rank:** Professor of Physical Chemistry

**Department and University:** Department of Chemistry  
Oklahoma State University  
Stillwater, OK 74078

**Research Location:** Frank J. Seiler Research Laboratory  
U.S. Air Force Academy  
Colorado Springs, CO 80840

**Focal Point:** Capt. Michael Coolidge

**Date:** July 23rd, 1991

## I. Abstract

The adducts of HF and HCl with aluminum hydrides, aluminum halides, aluminum hydrohalides, aluminum hydroxides and aluminum oxides have been studied using both semi-empirical and ab initio molecular orbital methods. When the electron rich region of HF or HCl is placed between 2.5 - 3.5 Å above the aluminum atom, Lewis Acid/Base adducts are generally found. In a few cases, the starting geometry rearranged to give a hydrogen bonded structure. In fewer cases, chemical reactions were observed. The structures observed are reported and briefly discussed.

## II. Introduction

Molecular Orbital Theory has been able to describe chemical bonding for several decades, first for pi-orbitals using Hückel Theory, then for valence orbitals using semi-empirical methods and finally, as computation power increased, for all orbitals by ab initio methods. For the past two decades, ab initio methods have been available and, in the main, the size of the system which could be studied was limited by computer power. Within the past five years the availability of supercomputers has permitted the study of large systems (more than 100 electrons) at a primitive level or the study of smaller systems in more sophisticated ways. Thus, one could study the van der Waal's interaction between two He atoms two decades ago but only within the past five years could weak interactions, say of the order of 3 to 30 kcal/mole, be studied and realistically compared with experiment. We can now handle very large systems, e. g. DNA, in a very crude way.

Weak interactions between molecules can be classified as van der Waal's type, in which no electrons are transferred (e.g. induced dipole-induced dipole or London Forces and dipole-dipole interactions, of which hydrogen bonds is the best example). A second type of weak interaction is the formation of Lewis Acid/Base adducts in which an electron rich region of one molecule can donate a small fraction of its electrons to an electron poor region of another. The classic intermolecular example is borane carbonyl,  $\text{H}_3\text{B}\cdots\text{CO}$ . The classic intramolecular example is diborane itself, where the empty p orbitals on the two borane molecules share the electrons involved in the B-H bond on the other borane molecule. This symmetric interaction is often called a three-center two-electron bond. Boric acid,  $\text{B}(\text{OH})_3$  is a monobasic acid because it does not donate protons but accepts  $\text{OH}^-$  and, since there is only one empty p orbital, the molecule is capable of neutralizing one  $\text{OH}^-$ . Clearly Lewis Acid/Base interactions, especially those that exceed thermal energy ( $3/2 \text{ RT}$ , 0.9 kcal/mole at room temperature) dictate the nature of the molecular entity that is present. This knowledge is fundamental to our understanding of the chemistry of these systems.

The author has become more and more involved in boron chemistry since 1985 when he first suggested that silylborane was involved in the deposition of p-doped a-silicon.<sup>1</sup> There followed a long series of papers dealing with fluorinated silyl boranes.<sup>2-4</sup> Recently, the author has shown that  $\text{BH}_3$  was capable of forming important adducts with  $^3\text{P}$  O atoms, with  $^3\Sigma$  dioxygen, and OH radicals.<sup>5</sup> These adducts may play an important and hitherto unsuspected roles in the oxidation of diborane.

Because borane, allane, gallane, etc. constitute higher energy fuels they are of interest to the Air Force, especially borane and allane. Having published exploratory calculations of the adducts of borane, it was natural for the author to expand these calculations to adducts formed by aluminum compounds. An impetus to do these calculations at the USAFA was provided because aluminum oxides constitute an important product in the exhaust of the solid state booster rockets for the Space Shuttle. Since HCl was also in the exhaust, it was only natural to study the adducts of aluminum compounds with HCl and, since HF is computationally more convenient, with HF as well. Also, although we did not appreciate this when we first came to the F. J. Seiler Laboratory, there was considerable interests in the application of aluminum halide salts as solvents for high energy electrochemical cells. These salts were eutectic mixes prepared by mixing  $\text{AlCl}_3$  with a  $\text{Cl}^-$  containing salt, such as NaCl or RCl, where R is an organic

quaternary ammonium salt. The reaction between  $\text{AlCl}_3$  and  $\text{Cl}^-$  is, of course, a Lewis Acid/Base reaction such as considered in this study. Thus, the proposed study of the HF and HCl adducts of aluminum hydrides, halides, hydroxides and oxides was of some interest to the Air Force.

### III. The Methods

Two molecular orbital methods are available for studying the the electronic properties of molecules, the semi-empirical method in which the two electron integrals are approximated using experimental data (parameterized), and the ab initio method in which nothing is parameterized. At the F. J. Seiler Laboratory, the only computers conveniently available were a cluster of  $\mu$ -Vaxes. These machines can be applied to ab initio calculations but they are so slow that such usage is not practical for the the majority of molecules studied here.  $\mu$ -Vaxes are much better suited to semi-empirical calculations and one of the best programs, MOPAC, was in fact developed at the Seiler Laboratory under the auspices of Dr. J. J. P. Stewart.

Before we fully appreciated the virtues (and limitations) of the MOPAC program we explored the adducts using the computationally "inexpensive" 3-21G\* basis set and the ab initio program. After optimization with this basis set, the molecules (adducts) were reoptimized using the 6-31G\* basis set. Then, at this level, correlation was taken into account using fourth order Møller-Plesset perturbation theory, MP4(SDTQ). Energies so determined are generally accepted to be within 1-2 kcal/mole of experimental values. Later, the 3-21G\* screening was abandoned in favor of MOPAC screening. We were sufficiently impressed with the speed and versatility of the MOPAC program that we plan to adopt it for all future preliminary screening.

It is important to note that all calculations were started with the electron rich end of the HF and HCl molecules in the empty p orbital, two-three angstroms above the plane of aluminum compounds and, hence, in the region where Lewis Acid/Base adducts were to be expected. The calculations were, therefore, biased toward the anticipated result. As will be seen, in most cases the expected adduct was found. However, there were a few interesting exceptions where the molecule rearranged/reacted rather than form an adduct. There was insufficient time and insufficient computer resources to explore each individual surface. Hence, we cannot claim that we have found the global minimum for each interaction. However, in every case we determined the vibrational spectrum of the stationary point reached to ensure that the programs had generated true minima and not transition states.

### IV. Division of the Problem

It was quickly apparent that a study of the HF and HCl adducts of  $\text{Al-H}$ (singlet/triplet),  $\text{AlH}_2$ ,  $\text{AlH}_3$ ,  $\text{Al-F}$ (singlet/triplet),  $\text{AlF}_2$ ,  $\text{AlF}_3$ ,  $\text{Al-Cl}$ (singlet/triplet),  $\text{AlCl}_2$ ,  $\text{AlCl}_3$ ,  $\text{AlH}_2\text{F}$ ,  $\text{AlHF}_2$ ,  $\text{AlH}_2\text{Cl}$ ,  $\text{AlHCl}_2$ ,  $\text{AlH}_2\text{OH}$ ,  $\text{AlH}(\text{OH})_2$ ,  $\text{Al}(\text{OH})_3$ ,  $\text{O=Al-H}$ ,  $\text{O=Al-F}$ ,  $\text{O=Al-Cl}$ ,  $\text{AlO}$ , and  $\text{Al}_2\text{O}_3$  was a rather prodigious undertaking. This required over 50 optimizations, frequency analyses, and MP4(SDTQ) calculations. [Some of these, e.g.,  $\text{HCl}\cdots\text{AlCl}_3$ , involved too many electrons (e.g. 82) to be performed using anything smaller and slower than an IBM 3090, such as located at Oklahoma State University, or the CRAY-2 computer, located at the National Center for Supercomputing Applications (NCSA), University of Illinois.] Therefore, we decided to break the calculations into two projects, one involving the HF and HCl adducts of aluminum hydrides and halides, the other the HF and HCl adducts of aluminum hydroxides and oxides. The manuscript for the first paper is being prepared concurrently with this report. The manuscript for the second paper will be finished early this Fall.

### V. Results

It would overly lengthen this report to discuss all of the results in hand. Hence, we shall summarize the observations here and refer the reader to the respective papers for additional details. Since

HF and HCl are dipolar and many of the fragments studied were polar, the geometry of the dipole-dipole interaction is not very different from that of the Lewis Acid/ Base adducts. However, one sure indication of electron donation (adduct formation) was the distortion of the aluminum compound from the  $sp^2$  hybridization (planar) toward the  $sp^3$  hybridization (tetrahedral). We show this generally in Figure 1.

In a few cases, usually with fragment aluminum compounds, the HF or HCl adduct rearranged to hydrogen bond with the negative region of the aluminum adduct, vide Figure 2, and the aluminum fragment hardly changed during the optimization process. The most remarkable of these structures was that for which  $X=H$ , i.e. the aluminum hydride fragment. In this case a hydrogen bond was formed between a hydride hydrogen and a proton-like hydrogen. To the author's knowledge, such a hydrogen bond has never been observed either theoretically or experimentally. While structures of this type were not observed for  $AlXYZ$  molecules, there is not a priori reason to rule them out. We feel that a thorough investigation of the surface of, say, HF and  $AlH_3$  would find such structures. As mentioned previously, our studies were biased toward Lewis Acid/Base Adducts and not toward the formation of hydrogen bonds. Starting geometries which favored hydrogen bond structures, i.e.,  $F-H \cdots H-AlH_2$ , would probably find such structures.

In a few cases, chemical reaction was found to occur. One example is the addition of HF to the aluminum oxygen double bond, illustrated in Figure 3. In addition we observed both hydrogen atom and halogen atom abstraction reactions.

For hydroxides and oxides both adduct formation and H-bonding was observed. Which process dominated and was found by optimization seemed to depend upon the nature of the acid (HF vs HCl) and on the nature of the aluminum compound. We illustrate the two possibilities in Figure 4. When X was a hydroxyl group, the hydrogen bond product was commonly formed.

The MP4(SDTQ) dissociation energies of the adducts ranged from 3.0 kcal/mole to 25.8 kcal/mol. In general the HCl adducts were more weakly bound, presumably because of their larger size. It would unduly lengthen this report to go into the variations in the dissociation energies observed. The reader is directed to the manuscripts for structural details, energies, and specific interpretations. Alternately, the reader can contact the author at Oklahoma State University.

## VI. Conclusions

- 1) The empty p orbitals in Group IIIA compounds leads to the formation of Lewis adducts between HCl and HF and the majority of the aluminum hydrides, halides, hydroxides and oxides.
- 2) In a few cases hydrogen bonding, i.e. dipole-dipole interaction, was observed. Since all the molecules studied were polar, one suspects that all of the aluminum compounds can form hydrogen bonds with HF and HCl. The reason these were not all observed in this study is believed to be a consequence of the initial starting geometry.
- 3) Based on these studies one can infer that HCl is certainly hydrogen bonded to the aluminum oxides/hydroxides in the exhaust of the Space Shuttle's solid state booster rockets. This is important since there is concern about the fate of HCl gas in the rocket exhaust.
- 4) HCl and HF should be highly soluble in  $AlCl_3$ -NaCl melts. These may offer a way of modifying the properties of these liquids. Similarly, aluminum hydroxide, chloroallane, etc. would be expected to dissolve in these melts to alter the liquid properties.

## VII. Future Work

Although we (the author and his Graduate Student, Mr. Marty Wilson) have conducted a broad study of the HF/HCl adducts of aluminum hydrides, halides, hydroxides and oxides, we have but barely scratched the surface of this field. Each of the 50 some adducts deserves a more detailed study and a full understanding of these interactions will only come from a careful exploration of each energy surface.

HF and HCl have scarcely exhausted the list of potential electron donors. One can think about water, ammonia, alcohols, amines, etc. as well. Further, since these adducts may not be the global minima one should search chemical reactions in which new chemical bonds were formed. For instance, in one case HF was found to add to react with  $\text{O}=\text{Al}-\text{OH}$  to produce  $\text{O}=\text{Al}-\text{F}$  and  $\text{H}_2\text{O}$ . Similarly, one must expect that  $\text{H}_2\text{O}$  reacts with  $\text{AlCl}_3$  to form  $\text{AlCl}_2\text{OH}$  and HCl, hydrogen bonded or as a Lewis adduct. The transition states for these reactions, none of which have been explored, certainly lie at shorter aluminum fluorine distances than assumed in the starting geometries used here. Much remains to be done before these systems are fully understood.

## References

1. G. J. Mains, C. Bock and M. Trachtman, A Theoretical Study of the Silyboranes, *J. Phys. Chem.*, **1985**, *89*, 2283.
2. G. J. Mains, C. Bock and M. Trachtman, A Theoretical Study of Difluorosilylborane, *J. Phys. Chem.*, **1986**, *90*, 51.
3. G. J. Mains, H. Niki, P. D. Maker, C. Bock and M. Trachtman, An Ab Initio Study of Silyldiborane, *J. Phys. Chem.*, **1986**, *90*, 5317.
4. G. J. Mains, C. Bock and M. Trachtman, Ab Initio Study II of Fluorinated Silylborane, *J. Phys. Chem.*, **1988**, *92*, 294.
5. G. J. Mains, Ab Initio Molecular Orbital Study of Adducts and Oxides of Boron Hydrides, *J. Phys. Chem.*, **1991**, *95*, 5089.

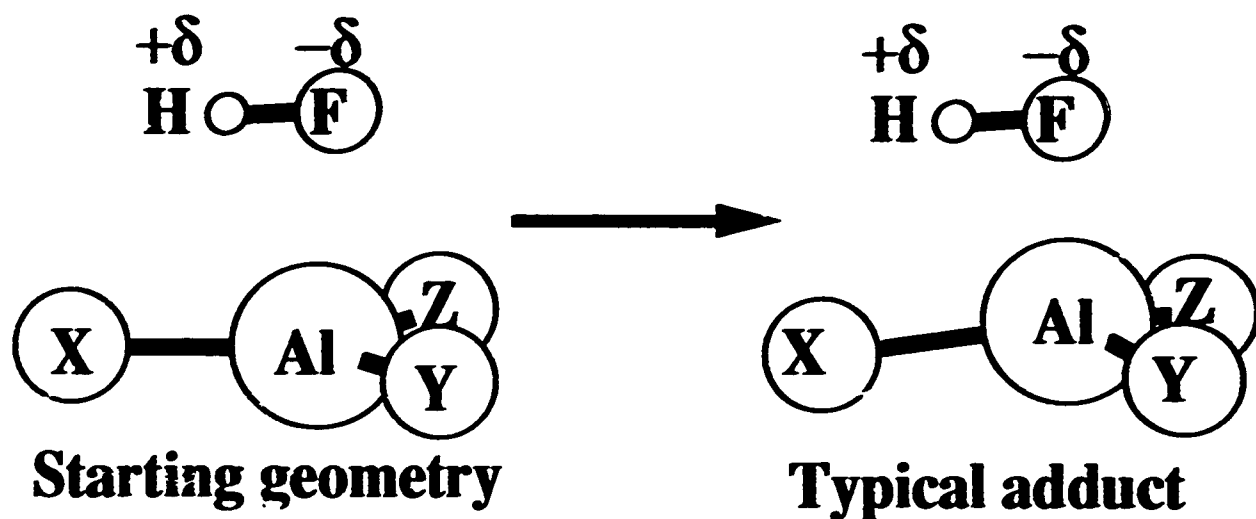


FIGURE 1.

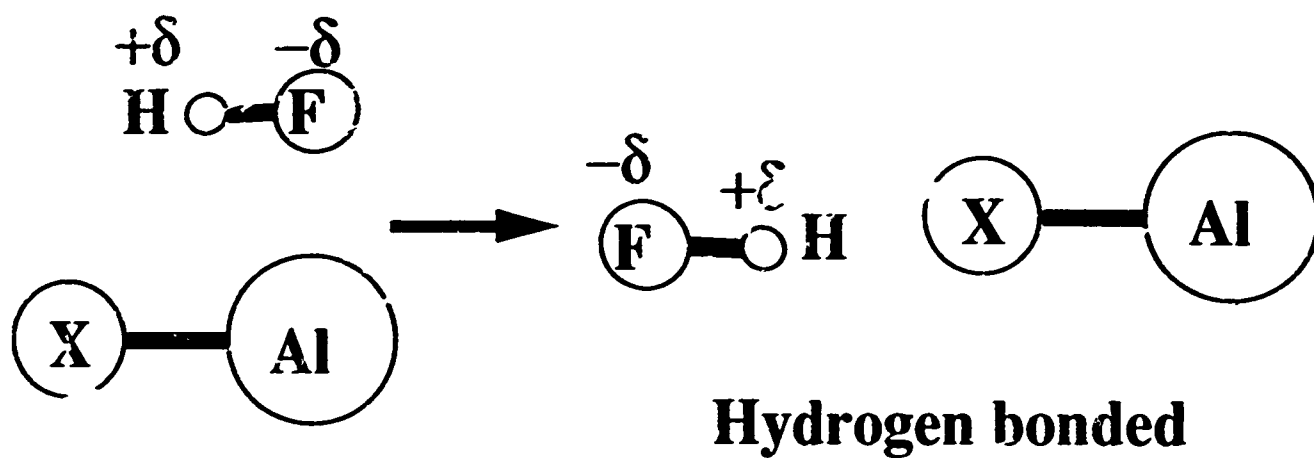
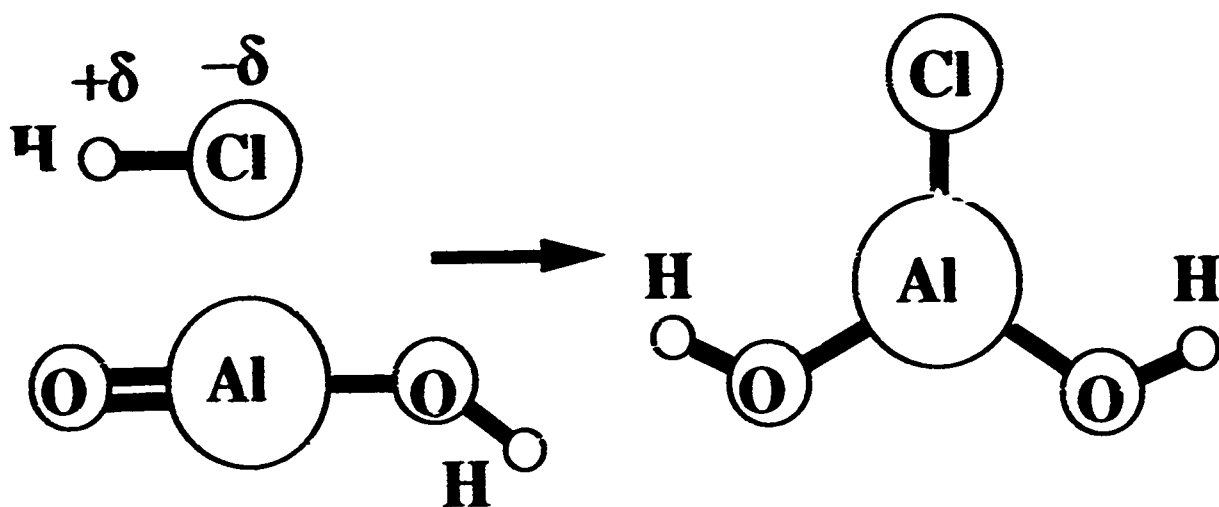
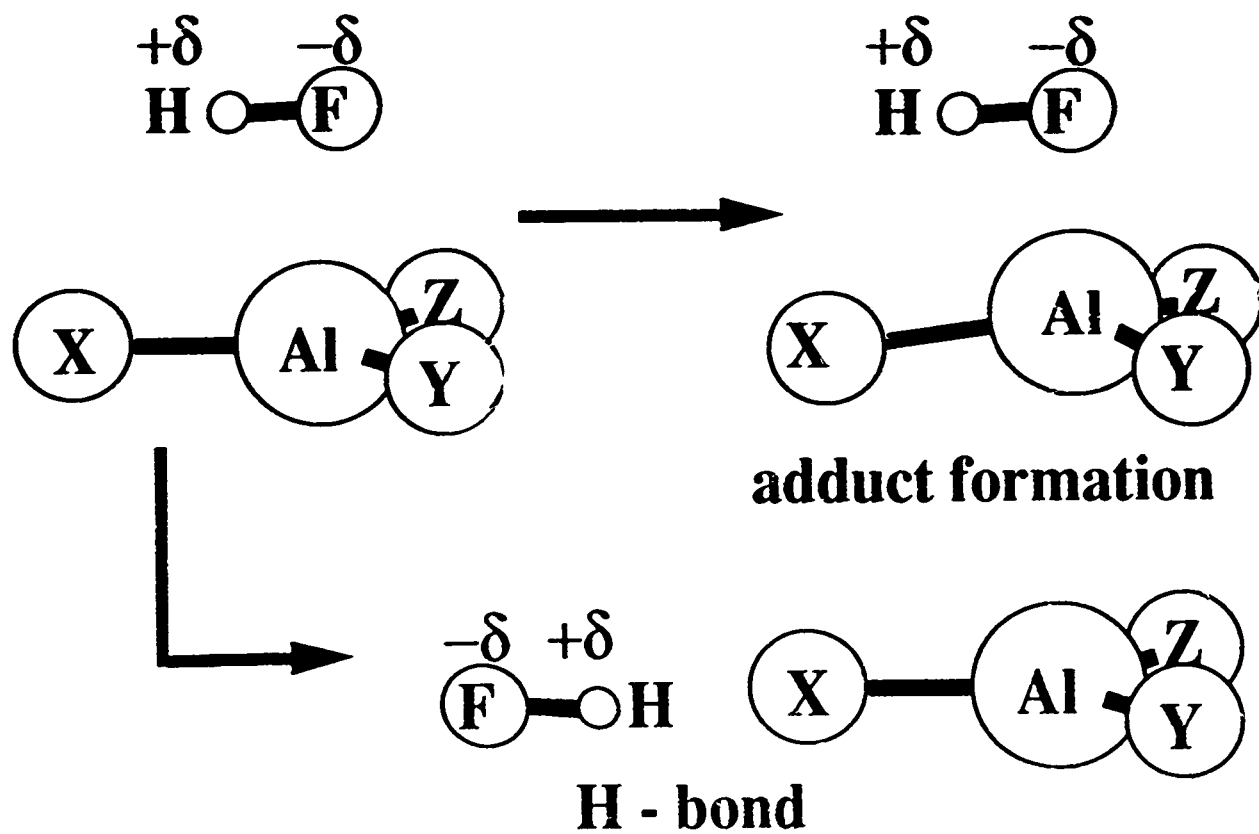


FIGURE 2





**FIGURE 3**



**FIGURE 4**